5-2017

Reflecting Human Knowledge of Place and Route-Choice Behavior Using Big Data

Jiaoli Chen
University of Tennessee, Knoxville, jchen42@vols.utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Jiaoli Chen entitled "Reflecting Human Knowledge of Place and Route-Choice Behavior Using Big Data." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Geography.

Shih-Lung Shaw, Major Professor

We have read this dissertation and recommend its acceptance:

Bruce A. Ralston, Hyun Kim, Lee D. Han

Accepted for the Council:
Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Reflecting Human Knowledge of Place and Route-Choice Behavior Using Big Data

A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Jiaoli Chen
May 2017

DEDICATION

I dedicate this dissertation to my parents, Huiyu Chen and Xiangping Li, who are always doing their best to love and support me. This dissertation is also dedicated to my fiancé, Ye Hao, who took care of me during my PhD research.

ACKNOWLEDGEMENTS

I am very grateful to my advisor, Dr. Shih-Lung Shaw. Dr. Shaw encouraged me to gain confidence in everything. It is a great wealth for both my career and life. His academic advice and support helped me get abilities of critical thinking and independent research. He also helped me realize the importance of real-world thinking. This is the biggest gain during my Ph.D. period and will definitely benefit my future career.

Many thanks to Dr. Bruce Ralston. From Dr. Ralston I have learnt the importance of keeping learning new things. He showed me what an active person should be like. He is always happy to share new technologies with students and willing to provide help.

I am very thankful to Dr. Hyun Kim and Dr. Lee Han. Dr. Kim is such a nice teacher and gave me many detailed suggestions for both course work and dissertation. Dr. Han also provided many useful suggestions for my dissertation research.

ABSTRACT

Exploring human knowledge of geographical space and related behavior not only helps in understanding human-environment interactions and dynamic geographic processes, but also advances Geographic Information Systems (GIS) toward a human-centric paradigm to make daily life more efficient. Today's relatively easy acquisition of various big data provides an unprecedented opportunity for geographers to answer research questions that previously could not be adequately addressed. However, new challenges also arise regarding data quality and bias as well as change in methodology for dealing with big data that are different from traditional data types.

Representing people's perception of place and studying driver's route-choice behavior are two of the many applications of big data in answering research questions about human knowledge and behavior in the fields of GIS and transportation. Incorporating three papers, this dissertation focuses on these two different applications to achieve the following objectives: 1) examine the degree to which a geographic place's spatial extent can be estimated from human-generated geotagged photos; 2) address the challenge of geotagged photos' uneven spatial distribution in place estimation and explore an approach that can better derive a place's spatial extent; 3) develop a method that can properly estimate the spatial extent of a place that has multiple disjoint regions while considering geotagged photos' uneven distribution; 4) explore useful spatiotemporal patterns of taxi drivers' route-choice behavior in a dynamic urban environment.

This dissertation makes three major contributions to big data applications' systematic theory: 1) proposes an effective approach to handling the uneven spatial distribution problem of geotagged photos as a type of volunteered geographic data by modeling their representativeness; 2) develops methods that can properly derive the vague spatial extent of a place with or without disjoint regions; and 3) explores taxi drivers' route-choice patterns in different situations that can inform future transportation decisions and policy-making processes.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Research Background

Human perception and knowledge of geographical space are often formed during people's interactions with environments by conducting various daily activities (Tuan 1977, Massey 1994). In turn, perception and knowledge influence human daily practice, such as inhabiting, traveling, communicating and working, which are the underlying contributors to human and urban dynamics. Acquiring information about human knowledge and behavior is important in studying such fields as Geographic Information System/Science (GIS) and transportation. This information can be used to achieve the following research goals: explaining geographic reality and process (Goodchild 2011), understanding human and urban dynamics, developing human-centric GIS tools and services, providing smarter navigation applications, mitigating urban traffic congestions, and designing efficient transportation and urban systems.

People's perception and knowledge of geographical space and related behavior are also revealed in a wide range of informal ways (e.g., use of place names to refer to geographic locations, use of words and sentences to describe one's feeling about and experience of the surroundings, pictures and videos of interest, and trajectories of one's route-choice decisions). Before the age of ubiquitous information and communications technology (ICT), data about such perception, knowledge and behavior was acquired through interviews, surveys and experiments. This kind of data collection is tedious, time-consuming and expensive. Furthermore, the data generally have low coverage of space and time, reflect a small proportion of the population, and are usually static. The challenge of obtaining large-scale datasets containing information about human perception and behavior once prevented researchers from answering interesting research questions about human dynamics.

In contrast, today's pervasive information and communication devices (e.g., computers, mobile phones, global positioning system (GPS), cameras and radio-frequency identification) record and track human discourses and activities with high spatial and temporal resolutions. The big data created in the digital world are now regarded to be valuable "exaflood" (Swanson 2007) in both business and scientific worlds (Sui et al. 2013). The GIS community has used those geographically related big data from different sources to address various geographic and transportation topics.

Different data types have varying degrees of suitability for examining different perspectives of human perception, knowledge and behavior. The major categories of most used geographically and/or temporally referenced big data and the related research topics regarding human perception, knowledge and behavior include the following:

1. Georeferenced social media data as sensors of geographic space's social content. These data are created by ordinary citizens during their participation in online social media, such as geotagged Flickr photos and tags, Twitter tweets, Facebook postings and Foursquare check-ins. Closer to human discourse than other data types containing only geolocations, they are often used to examine geographic space's social side, such as harvesting human perception of a place (e.g., Gao et al. 2014) and inferring social characteristics about physical space (e.g., Graham and Zook 2011, McKenzie et al. 2015). Researchers also use these data to explore human behavior and activity patterns (e.g., Azmandian et al. 2013, Hasan et al. 2013).

2. Tracking data collected from GPS devices. These trajectory data have been applied to explore and understand the spatiotemporal patterns of human activity-travel behavior as well as resulting human and urban dynamics, such as dynamic urban flows (e.g., Giannotti et al. 2011, Veloso et al. 2011), city hotspot extraction (e.g., Palma et al. 2008), collective human mobility (e.g., Liu et al. 2012, Jiang et al. 2009), time-varying traffic condition (Ehmke et al. 2012), and routing-strategy differences between drivers (Liu et al. 2010).

3. Location data collected from mobile phone devices that are more pervasive provide individual persons' digital footprints. The unit of cell phone data (individual person) is "finer" (i.e., more direct human footprint at the individual level) than that of vehicle GPS tracking data (vehicle). Thus, such data are frequently used to study human-activity space (Yuan et al. 2012, Xu et al. 2016) and mobility patterns (González et al. 2008).

These big data bring not only opportunities for studying human dynamics but also new challenges when dealing with the issues caused by the three V's of big data: volume, velocity and variety (Laney 2001). Thus, new methods and tools are needed to increase capabilities to process, analyze and visualize such new type of data (Sui and Goodchild 2011). Miller and Goodchild (2015) suggest that in data-driven science, GIS methods and tools should shift from finding explanations and universal laws to exploring specific patterns and descriptions at certain places and times. Methods should also be specifically developed for certain data types and research questions.

Addressing specific opportunities and challenges for different types of big data and research questions, this dissertation focuses on two of the many perspectives in big data applications: reflecting human perception of place and understanding route-choice behavior. The concepts of place and route-choice behavior play important roles in human dynamics and are examined in this dissertation on the scope of human dynamics. Human perception of place (e.g., home, work place, business place and commercial place) greatly influences urban system and travel demands, which are directly related to human dynamics in a city. Route-choice

decisions between two places can both influence and be influenced by time-varying flows in urban dynamics. More details regarding specific research questions are discussed in the following subsections.

1.1.1 Reflecting Human Perception of Place Using Geotagged Photos

The concept of place comes from associating social meanings to physical space through human interactions and experiences with geographic environments (Tuan 1977, Agnew 2011). This concept is common in human discourse (e.g., using place names in daily conversations to specify locations). People often ask such questions as "where is Chinatown in New York City," but its extent usually cannot be located on a map. *Chinatown* is a vernacular name coming from people's understanding of a neighborhood's social characteristics but has not been officially defined by authorities. Thus, it is a vague concept in terms of undefined geographic boundary and varying perceptions among people. GIS tools and services need to better represent place from a spatial perspective to bridge informal human discourse and a precise computational environment.

Human perception and knowledge of a place are difficult to identify until they are explicitly expressed. During the past decade, GIS and social media have been converging (Sui and Goodchild 2011). For example, most social media websites provide a location-sharing function enabling users to assign a geolocation to their postings (e.g., Twitter tweets, Facebook status updates, and Flickr photos). From the geographic perspective, this geotagging is a process in which people explicitly express their ideas about a location based on their experience, knowledge, understanding and feeling. Among the geotagged textual contents, place names are included and their relationship to corresponding geolocations are established. Thus, a place's spatial extent can be estimated using the geolocations that social media users consider to be inside the place.

The convergence of GIS and social media has brought an era of humans as sensors (Goodchild 2007) in which a bottom-up crowdsourcing process of average citizens' geographic data production is emerging to supplement the traditional top-down authoritative process of geographic data production (Sui et al. 2013). This approach promotes interest in researching human knowledge of place based on volunteered geographic data that are unprecedentedly large scale and dynamic. However, the three V's issues associated with big data also exist in volunteered geographic information (VGI) (Goodchild et al. 2016), and new methodologies are needed. Specifically, when such VGI is used to reflect human perception and knowledge of place, challenges exist in terms of systematically biased representation in user groups and spatial coverage (Lüscher and Weibel 2013, Hollenstein and Purves 2010) as well as lack of quality assurance (Goodchild and Li 2012). In other words, social media users who contribute to the VGI content may not represent the perception and knowledge of people who do not use social media.

Moreover, geotagged content's availability and distribution are inconsistent over the entire geographic space. For example, less or no content is available in unpopular areas. Furthermore, VGI's contributors are average citizens who might not be well trained with geographic knowledge. They tend to make mistakes and reduce VGI's quality.

Given the above challenges in using human-generated content conveying information about human perception and knowledge of a place to explicitly represent and visualize such information, this dissertation aims to address the following research questions:

> To what degree can biased representation and no quality assurance influence the estimation of place extents based on geotagged photos? Can these problems be overcome? What are some effective approaches to overcoming them?
>
> How effective are the geotagged photos for deriving vague spatial extents of places? To what degree can these extents be correctly approximated?

1.1.2 Exploring Spatiotemporal Patterns of Route-Choice Behavior Using GPS Tracking Data

Reflecting human perception and knowledge of place discussed in Subsection 1.1.1 advances GIS tools and services toward a human-centric paradigm by helping answer a real-world question about a place's location. Acquiring route-choice behavior's spatiotemporal patterns can also help improve GIS services' performance in navigation applications by answering a real-world question about what is a good route between two places.

The mutual influence between human route-choice behavior and a transportation system is one of urban dynamics' major components. Route-choice decisions are made under various constraints, such as space and time, and can influence the level of human mobility. Development of transportation systems depends heavily on the characteristics of urban residents' activity-travel behaviors, which include route-choice behavior. After Hägerstrand (1970) initiated the era of *time geography*, transportation research's focus changed from conventional trip-based aggregate approaches to activity-based disaggregate approaches under space-time constraints (Timmermans et al. 2002). This focus requires more detailed information about route-choice behavior at the individual level and under different space-time situations. The commonly used assumption of optimal route could be no longer applicable if urban dynamics and human factors are considered (Li et al. 2011). Although the literature shows great efforts in examining the factors influencing route-choice decisions (e.g., Hölscher et al. 2011, Papinski et al. 2009),

route-choice behavior's dynamic property and spatial variation have not often been considered.

A city's taxi GPS tracking data include detailed information about which routes taxi drivers chose and what strategies were used to avoid bad traffic conditions (Castro et al. 2013). This large dataset can help explore spatiotemporal patterns of route-choice behavior and validate conclusions from existing studies. However, mining such big trajectory data involves some challenges: large data size, multiple dimensions, and coexistence of homogeneity and heterogeneity in human behavior patterns. Existing data-mining and knowledge-discovery approaches have difficulty dealing with space-time constraints in human movements. These challenges call for higher computational power and innovative spatiotemporal data-mining approaches while simultaneously considering space and time. Thus, this dissertation focuses on developing effective and efficient space-time approaches to exploring spatiotemporal patterns of taxi drivers' route-choice behavior considering urban dynamics. Specific research questions include the following:

> Where, when, and to what extent do taxi drivers deviate from the shortest-distance routes? What are some effective methods for facilitating spatiotemporal analysis of taxi drivers' deviation patterns?
>
> Are road functional class, travel distance and urban rush hours related to taxi drivers' deviations from the shortest-distance routes? Which factors are more influential? What unknown patterns can be found from big taxi tracking data? Are the conclusions drawn from such big data consistent with those from traditional survey and sample data?

1.2 Organization of Chapters

The remainder of this dissertation consists of three paper chapters plus one concluding chapter. Each paper chapter is written in the form of journal article. Chapters 2 and 3 answer the research questions listed in Subsection 1.1.1 from the perspective of human perception and knowledge of place. Chapter 4 answers the research questions in Subsection 1.1.2 from the perspective of human route-choice behavior. These three chapters are organized as three independent papers.

Chapter 2 explores the important impact of geotagged photos' uneven spatial distribution on estimating vague place extents. To overcome the biased spatial coverage, an approach is proposed for modeling the representativeness of each geotagged photo point based on its location popularity. A modified kernel density estimation method incorporating photo representativeness is developed. It is tested with eight places, which cover urban vs. non-urban areas, with vs. without

an official boundary cases, and at various spatial scales of state, city and district levels. The test results indicate noticeable improvements of the proposed representativeness-weighted kernel density estimation (RW-KDE) method over the traditional kernel density estimation method in estimating places' vague spatial extent. The results also indicate that using geotagged photos is feasible to derive acceptable spatial extent of places once the unevenly distributed photos' representativeness is properly adjusted.

Chapter 3 identifies a challenge caused by outlier photos (i.e., photos tagged with a target place name but not located within the target place) when using geotagged photos to estimate the spatial extent of a place with multiple disjoint extents located in different regions of a large study area. This chapter proposes a method named *representativeness-weighted kernel density estimation for disjoint extents* (RW-KDE-FDE). This method first uses a scan statistic approach to determine the number and rough locations of a place's disjoint extents, and then applies an improved outlier-removal process and RW-KDE approach to derive the vague extents. The method is tested with three places that have disjoint extents in study areas of different scales. The results show a place's disjoint extents and their improved spatial representation.

Chapter 4 focuses on taxi drivers' route-choice patterns at different locations and times and in different situations by comparing their actual routes to the shortest-distance routes. Two indices measuring taxi drivers' preference for and avoidance of each road segment are proposed to help reveal the spatiotemporal patterns of taxi drivers' deviations from the shortest-distance routes. This chapter finds that deviation from the shortest-distance route is influenced more by road functional class and travel distance than by urban rush hours. Taxi drivers tend to frequently detour from the shortest-distance routes in areas with a high density of local streets, but tend to follow the shortest-distance routes on high-hierarchy roads and short trips. Taxi drivers are not likely to choose a primary road as an alternative if it is not on the shortest-distance route, but once a primary road is on the shortest-distance route, they tend to stay on it. By examining the difference in deviation rate between rush hours and off-rush hours, this chapter finds that the aggregate patterns of taxi drivers' deviations are relatively stable regardless of urban rush hours.

Chapter 5 summarizes this dissertation research's major contributions and identifies limitations as well as future research directions.

References

Agnew, J.A., 2011. Space and place. *In*: J.A. Agnew and D.N. Livingstone, eds. *The SAGE handbook of geographical knowledge*. Thousand Oaks, CA: SAGE, 316–330.

Azmandian, M., et al., 2013. Following human mobility using tweets. *In*: L. Cao, et al., eds. *Agents and data mining interaction. ADMI 2012. Lecture notes in computer science vol. 7607*. Berlin: Springer, 139-149.

Ehmke, J. F., Meisel, S., and Mattfeld, D. C., 2012. Floating car based travel times for city logistics. *Transportation Research Part C: Emerging Technologies*, 21(1), 338-352.

Gao, S., et al., 2014. Constructing gazetteers from volunteered big geo-data based on Hadoop. *Computers, Environment and Urban Systems*.

Castro, P. S., et al., 2013. From taxi GPS traces to social and community dynamics. *ACM Computing Surveys*, 46(2), 1-34.

González, M.C., Hidalgo, C.A., and Barabási, A.L., 2008. Understanding individual human mobility patterns. *Nature*, 453(7196), 779–782.

Goodchild, M.F., 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4), 211-221.

Goodchild, M.F., 2011. Formalizing place in geographic information systems. *In*: L.M. Burton, et al., eds. *Communities, neighborhoods, and health*. New York: Springer, 21-33.

Goodchild, M.F., Aubrecht, C., and Bhaduri, B., 2016. New questions and a changing focus in advanced VGI research. *Transactions in GIS*, 1-2.

Goodchild, M.F. and Li, L., 2012. Assuring the quality of volunteered geographic information. *Spatial Statistics*, 1, 110-120.

Giannotti, F., et al., 2011. Unveiling the complexity of human mobility by querying and mining massive trajectory data. *The VLDB Journal—the International Journal on Very Large Data Bases*, 20(5), 695-719.

Hägerstrand, T., 1970. What about people in regional science? *Papers in Regional Science*, 24(1), 7-24.

Hasan, S., Zhan, X., and Ukkusuri, S.V., 2013. Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. *In*: *Proceedings of the 2nd ACM SIGKDD international workshop on urban computing*. New York: ACM.

Hollenstein, L. and Purves, R., 2010. Exploring place through user-generated content: using Flickr tags to describe city cores. *Journal of Spatial Information Science*, 1, 21-48.

Li, Q., et al., 2011. Path-finding through flexible hierarchical road networks: An experiential approach using taxi trajectory data. *International Journal of Applied Earth Observation and Geoinformation*, 13(1), 110-119.

Hölscher, C., Tenbrink, T., and Wiener, J. M, 2011. Would you follow your own route description? Cognitive strategies in urban route planning. *Cognition*, 121(2), 228-247.

Lüscher, P. and Weibel, R., 2013. Exploiting empirical knowledge for automatic delineation of city centres from large-scale topographic databases. *Computers, Environment and Urban Systems*, 37, 18-34.

Massey, D., 1994. Space, place and gender. Cambridge: Polity Press.

McKenzie, G., et al., 2015. How where is when? On the regional variability and resolution of geosocial temporal signatures for points of interest. *Computers, Environment and Urban Systems*, 54, 336-346.

Miller, H.J. and Goodchild, M.F., 2015. Data-driven geography. *GeoJournal*, 80(4), 449-461.

Graham, M. and Zook, M., 2011. Visualizing global cyberscapes: Mapping user-generated placemarks. *Journal of Urban Technology*, 18(1), 115-132.

Jiang, B., Yin, J., and Zhao, S., 2009. Characterizing the human mobility pattern in a large street network. *Physical Review E*, 80(2), 021136.

Laney, D., 2001. *3D data management: Controlling data volume, velocity, and variety*. META Group. Available from: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf [Accessed November 1, 2016].

Liu, L., Andris, C., and Ratti, C., 2010. Uncovering cabdrivers' behavior patterns from their digital traces. *Computers, Environment and Urban Systems*, 34(6), 541-548.

Liu, Y., et al., 2012. Understanding intra-urban trip patterns from taxi trajectory data. *Journal of Geographical Systems*, 14(4), 463-483.

Palma, A. T., et al., 2008. A clustering-based approach for discovering interesting places in trajectories. *In*: *Proceedings of the 2008 ACM symposium on applied computing*. New York: ACM, 863-868.

Papinski, D., Scott, D. M., and Doherty, S. T., 2009. Exploring the route choice decision-making process: A comparison of planned and observed routes obtained using person-based GPS. *Transportation Research Part F: Traffic Psychology and Behaviour*, 12(4), 347-358.

Sui, D. and Goodchild, M., 2011. The convergence of GIS and social media: challenges for GIScience. *International Journal of Geographical Information Science*, 25(11), 1737-1748.

Sui, D., Goodchild, M., and Elwood, S., 2013. Volunteered geographic information, the exaflood, and the growing digital divide. *In*: D. Sui, S. Elwood, and M. Goodchild, eds. *Crowdsourcing geographic knowledge*. Netherlands: Springer, 1-12.

Swanson, B., 2007. *The coming exaflood*. The Wall Street Journal. Available from: http://www.wsj.com/articles/SB116925820512582318 [Accessed October 15, 2016].

Tuan, Y.F., 1977. *Space and place: the perspective of experience*. Minneapolis: Uiversity of Minnesota Press.

Veloso, M., Phithakkitnukoon, S., and Bento, C., 2011. Urban mobility study using taxi traces. In: *Proceedings of the 2011 international workshop on trajectory data mining and analysis*. New York: ACM, 23-30.

Timmermans, H., Arentze, T., and Joh, C.H., 2002. Analysing space-time behaviour: new approaches to old problems. *Progress in Human Geography*, 26(2), 175-190.

Xu, Y., et al., 2016. Another tale of two cities: understanding human activity space using actively tracked cellphone location data. *Annals of the American Association of Geographers*, 106(2), 489-502.

Yuan Y, Raubal, M., and Liu, Y., 2012. Correlating mobile phone usage and travel behavior - A case study of Harbin, China. *Computers, Environment and Urban Systems*, 36(2), 118–130.

Chapter 2

Representing the Spatial Extent of Places Based on Flickr Photos with a Representativeness-Weighted Kernel Density Estimation

This chapter has been published as: Chen, J. and Shaw, S.L., 2016. Representing the spatial extent of places based on Flickr photos with a representativeness-weighted kernel density estimation. *In*: J.A. Miller, D. O'Sullivan, and N. Wiegand, eds. *Geographic Information Science. Lecture notes in computer science vol. 9927*. Cham: Springer, 130-144.

The research work was completed by Jiaoli Chen advised by Dr. Shih-Lung Shaw.

Abstract

Geotagged photos have been applied by many researchers to explore the spatial extent of places. This paper addresses an important challenge of using geotagged Flickr photos to delineate the spatial extent of a vague place, which is defined as a place without a clearly defined boundary. We argue that the variation of location popularity has a great impact on the estimation of such vague spatial extent of a place. We propose an approach to model the representativeness of each geotagged photo point based on its location popularity. A modified kernel density estimation method incorporating the photo representativeness is developed and tested with eight places, which cover urban vs. non-urban areas, with vs. without an official boundary cases, and at various spatial scales of state, city and district levels. Our results indicate major improvements of the proposed representativeness-weighted kernel density estimation method over the traditional kernel density estimation method in estimating the spatial extent of vague places.

2.1 Introduction

Naïve geography (Egenhofer and Mark 1995) envisions that the advanced geographic information systems (GIS) should "follow human intuition" (p. 1) and "support common-sense reasoning" (p. 5) so that ordinary people who do not need to know about GIS can use them easily. It is important for such GIS to understand and represent linguistic place names. Some places (e.g., administrative divisions) have a formally defined geographic extent to be represented in GIS, while many places (e.g., vernacular places) have no formally defined boundary but a vague geographic extent. Therefore, an effective and efficient representation of vague spatial extent of places is critical to GIS representation, query, analysis, and visualization of places.

Acquiring human knowledge of places is a traditional way to derive vague place extents. With the increasing popularity of geotagged social media (e.g., Flickr, Twitter, and Facebook), large numbers of place names exist in the contents from such platforms, carrying valuable information about people's perception of places. Among these social media, Flickr provides adequate and more direct associations between photo geolocations and place name tags (Martins 2011). Thus, extracting the spatial extent of places from Flickr data has been a GIS research topic (e.g.,

Grothe and Schaab 2009, Hollenstein and Purves 2010, Li and Goodchild 2012, Cunha and Martins 2014), especially for purposes of enriching large-scale gazetteers and GIS services (e.g., Keßler et al. 2009a, Gao et al. 2014). In the meantime, such crowd-sourced data also present challenges to place-related research, including to which degree the spatial extent of a place can be estimated from geotagged photos or other crowd-sourced social media data.

Past research has indicated the effectiveness of using Flickr data to identify major locations of vague places (e.g., Hollenstein and Purves 2010, Keßler et al. 2009b). However, estimation of vague geographic extents still presents a major challenge. As suggested by Jones et al. (2008), we argue that the underlying assumption of a random distribution of Flicker photos is incorrect when applying geotagged photos to estimate vague place extents. In general, there are fewer photos taken at locations with low accessibility and low popularity than those popular and easily accessible locations. The conventional approach of treating each geotagged photo with equal representativeness or importance, regardless of where it is located, is questionable. The representativeness of photo points located in unpopular areas could be under-weighted due to a low absolute number of photos taken in such areas, while the representativeness of photo points located in popular areas would be over-weighted due to a larger number of photos taken in these areas. As a result, unpopular locations (e.g., inaccessible parts of mountain areas) of a vague place (e.g., Rocky Mountains) would be significantly underestimated or even excluded in the derived geographic extent. On the other hand, popular locations could be overestimated and distort the boundary of a place. When the kernel density estimation (KDE) method is applied to delineate vague boundaries (e.g., Li and Goodchild 2012), the resulting surface usually looks like a hot spot map of the photos tagged with a target place name rather than a “probability field” (p. 205) (Goodchild et al. 1998) representing the place extent where a higher estimate indicates a higher probability of belonging to a target place. Thus, this study aims at improving the representation of vague place extents by adjusting photo point representativeness based on their location popularity. Note that we adopt the same term “target place” (p.1047) from Jones et al. (2008) to refer to a place whose extent needs to be estimated.

In the remainder of this paper, we start with a review of work related to the concept of place, georeferencing place names, and delineation of vague place extent from survey, web and social media data. In the methodology part, we discuss a proposed approach based on photo point representativeness and the representativeness-weighted KDE (RW-KDE) method. Next, we present the results of testing our proposed assumptions and method based on eight selected sample places, followed by comparisons with the results derived from the traditional KDE approach. We conclude this paper with contributions, limitations and future research directions.

2.2 Related Work

As an important concept in geography, place has been extensively studied, implying more than space by incorporating social, economic, cultural and political meanings through human experience (Relph 1976, Tuan 1977, Agnew 2011). Places are usually revealed in unstructured forms. Their informality is also reflected by the variations in people's understanding of them (Twaroch et al. 2009). Thus, they are often absent from current GIS in which geographic objects need to be disambiguated, abstracted and digitalized from the perspective of space. Great efforts have been made toward the convergence of place and GIS, such as theoretically modeling place in a computational environment (e.g., Cohn and Gotts 1996, Dilo et al. 2007) and delineating the spatial extent of place names which is the focus of this paper.

There have been some place-friendly web applications using simple gazetteers to handle place names (Jones et al. 2008). But retrieval of vague place extent is still beyond the ability of current gazetteers. Usually given an authority-recognized place name, a gazetteer provides information about its geographic location, feature type, and relationship to other places (Goodchild and Hill 2008). But it provides limited information on vernacular names. In most gazetteers, the location of a place with a large spatial extent is usually represented as a point, rectangle bounding box, or occasionally a polygon with crisp boundary (Twaroch et al. 2009). Although an abstract and simple geometry benefits computational process in current information systems, it falls short in delivering information about the inherent vagueness of place boundary which often reflects human perception and cognition. Therefore, much research has focused on deriving vague place extents from a variety of data sources.

Montello et al. (2003) conducted an empirical study in which human subjects were asked to draw shapes of the downtown Santa Barbara area based on their understanding of the place extent. The downtown shapes drawn by different participants were then aggregated to generate a probabilistic representation of the vague place extent. Montello et al. (2014) interviewed another two groups of participants to unveil and measure the variation of vagueness of a place perceived by different people and at different locations. These survey-based approaches have an advantage that data structure and collection procedure can be designed to facilitate subsequent modeling and estimation of vague place extents as well as to answer specific research questions. However, the difficulty in collecting such empirical data obstructs their wide applications.

Some other research derived a place's extent using its topological relations to other clearly defined geographic objects. Based on a set of points covering a region, Parker and Downs (2013) combined DBSCAN clustering technique and fuzzy set theory for the delineation of a vague extent. For another example, based on two sets of geographic points that fall inside and outside a place, Alani et al.

(2001) created a Voronoi diagram of these points and delineated the place extent from the Voronoi polygons. Their method only generated a crisp boundary. Taking advantage of a diversity of spatial information, Schockaert et al. (2011) proposed a unique approach which could derive constraints from qualitative and quantitative spatial data and approximate a vague extent based on the derived constraints using techniques of genetic algorithm and ant colony optimization. All these approaches require that the geographic objects used as references are available and their geometries are already defined. They do not focus on how and where to collect these references and their spatial relations to the target place. For a large number of vernacular places, the practicality of these approaches depends on the availability of well-defined reference data.

Since a lot of place-related data can now be found from the web (e.g., web articles and documents, Internet yellow pages), much research used search engines to acquire references (e.g., hotels, cities) that are related to (e.g., inside, outside, covering, containing same words as) the target place name (e.g., Purves et al. 2005, Schockaert et al. 2005, Arampatzis et al. 2006). After the geolocations of returned references were determined by geoparsing and georeferencing, the vague extents were then estimated from the reference locations using KDE approach (Jones et al. 2008, Twaroch et al. 2009), fuzzy-set approach (Schockaert et al. 2005), or adapted α-shape and recoloring algorithms (Arampatzis et al. 2006).

As online social media became popular, geotagged photos (often from Flickr) were widely used in recent research for place extent assessment and digital gazetteer enrichment. This is because, unlike other web-sourced data, they do not require the geoparsing and geocoding processes which could introduce unexpected errors (Martins 2011, Grothe and Schaab 2009). Li and Goodchild (2012) applied KDE to Flickr geotagged photos, and found that the highest-density cells normally tell the major location of a place. They also pointed out a limitation of the data source being lack of sampling strategy. Martins (2011) improved the KDE surface by removing overestimated locations based on land coverage information, assuming that a place boundary is usually related to a land cover change. But this study did not address underestimated locations. Instead of the local density perspective in traditional KDE, Grothe and Schaab (2009) took a global perspective to generate the crisp boundary of places using a support vector machine (SVM) classification technique. Cunha and Martins (2014) improved the SVM method by incorporating place semantics from Flickr photo tags, demographic characteristic from population dataset, and topographical characteristic from elevation and land cover datasets, considering that a place's boundary can be found along the line where these characteristics change. The boundary they generated was also crisp and did not handle vague place extent. There is limited research in the literature that both delineates the vagueness of place extents and deals with the variation of location

popularity. Given this challenge, this study focuses on improving estimation of vague place extents based on geotagged Flickr photos.

KDE is a widely adopted method for estimating geographic extent in various domains such as animal home range (Downs and Horner 2012). It is relatively easy to use, fits well with data such as geotagged photos, and thus is frequently adopted by researchers. KDE generates a density surface through interpolation from the geographic points covering a place to reflect the inherent vagueness of place extent (Jones et al. 2008). Unlike fuzzy-set methods, KDE avoids using a subjective fuzzy membership function to represent vagueness. In the following sections, a modified KDE approach along with its assumptions are discussed and it is evaluated with data of eight selected case studies.

2.3 Methodology

2.3.1 Data Acquisition and Preprocessing

In order to assess if the performance of our proposed method works for different place types and at various feature scales, we selected the following eight places as case studies: *Manhattan Chinatown* and *San Francisco Chinatown* that do not have an official boundary at the urban district scale; City of *Nashville* and City of *Philadelphia* with an official boundary at the urban city scale; *Rocky Mountain National Park* and *Great Smoky Mountains National Park* as non-urban features with an official boundary; State of *California* and State of *Utah* with an official boundary at the state scale. These places are denoted as the *target place*s.

Around each target place, a larger rectangle study area (about eight times larger in size) was defined and used to search for the data of all geotagged photos through the Flickr search API. Note that we downloaded the data of all geotagged photos, not just those tagged with the target place name. In the reminder of this paper, we use the term *target photos* to refer to the data of photos tagged with the target place name or its variants, to distinguish them from the data of the whole set of photos denoted as *all photos*. The time span used in our searches varies among the eight places. For Manhattan Chinatown and San Francisco Chinatown, we searched for photos between January 2013 and February 2015 since these two places did not have an official boundary and their geographic extents could change over time. For the validation purpose, we extended Liu's (2014) approach of using the Street View imagery of Google Maps (https://www.google.com/maps) to manually delineate the boundary lines that separate locations with typical Chinese characteristics from those without Chinese themes. The derived boundary lines served as the benchmark reference to be compared with the geographic extents estimated by our proposed method. Most of the Chinatown Street View images were captured after 2013, which should be comparable with the time period of Flickr photos used in this study. For California, we downloaded Flickr photos

posted in January and July of 2014 as our sample data to test if partial Flickr data could also give acceptable estimates. Similarly, for Utah, only photos between September and December of 2014 were used. The time span chosen for the other four selected case-study places was between February 2004 and February 2015 since they all had an official boundary that were relatively stable over time.

As suggested by Liu (2014), redundant photos that were uploaded in bulk by the same user at the same location were removed to minimize the bias, with only one randomly chosen photo kept; place name variants (e.g., California abbreviated as CA; Nashville misspelled as Nashvile; Manhattan Chinatown shorten to Chinatown when the photo is posted on the Manhattan Island) were found by browsing through tags that occurred more than three times in the set of all photos. We then selected a set of target photos that were tagged with the place name or any acceptable variant from the set of all photos. Finally, based on the geolocations of target photos and all photos, two sets of geographic points for each place name were created respectively: *target points* and *all points*.

2.3.2 Representativeness of Geotagged Photos

The target points are those points assumed to fall inside the target place and used in many published works to generate the KDE surface. Here we denote the set of target points by T, and the set of all points by A. Their relation can be depicted by $T \subseteq A$. As discussed in the introduction section, unpopular locations tend to be underestimated, while popular locations tend to be overestimated. Thus, the contribution of a target point at a low popularity location should be increased to compensate for the disadvantage of photo availability at that location. Flickr data provide an opportunity to measure a location's popularity in terms of the photo volume. This makes it possible to model a target point's contribution level, namely representativeness, of the place in a study. Therefore, we assume that the location popularity is indicated by the volume of all points at a given location.

For implementation, we first discretize the earth surface in each study area into a regular grid with a cell size of $n \times n$, where n is the width of a cell. Each grid cell represents a location l. The popularity p_l of location l can be quantified by the count of all points falling in that cell. As stated above, when a location's popularity decreases, the representativeness of a target point at that location should increase. We assume that the representativeness r_t of target point t that falls in the cell of location k is inversely proportional to location k's popularity p_k. That is,

$$r_t = \frac{1}{p_k} \tag{2.1}$$

The function creates the value of representativeness falling within (0, 1]. The choice of grid cell size n is an important step. It can be observed from the set of all points in Figure 2.1 that the sparsest points located in the most unpopular locations

are usually isolated and distant from their nearest neighbor. The most isolated target point that has the largest nearest neighbor distance can be found and denoted as the *sparsest target point* which can reflect the most unpopular location near target points. Note that the nearest neighbor of a target point here is based on the set of all points. In order to maximize the possibility that the sparsest target point has the highest representativeness value 1 and to avoid assigning too many other target points with the value 1, we define *n* as the sparsest target point's nearest neighbor distance *d* divided by √2, which is:

$$d = max_{t \in T}\left(min_{a \in A-\{t\}} dist(t,a)\right);\ n = d/\sqrt{2} \tag{2.2}$$

where *t* represents a point in the set of target points *T*, and *a* represents a point in the set of all points *A*.

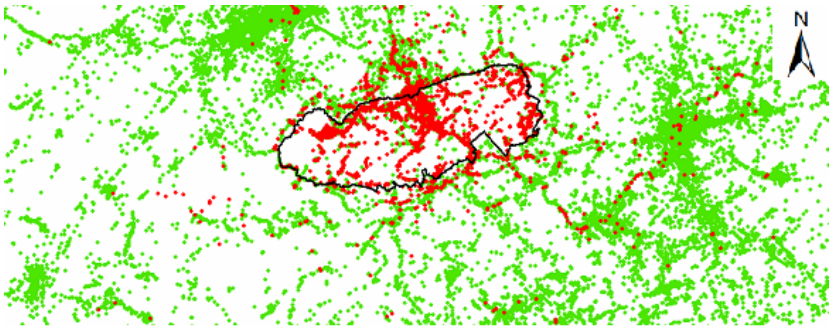


Figure 2.1 Spatial distributions of target points (red) and all points (red and green) of the Great Smoky Mountains National Park. The official boundary is shown in black solid line.

2.3.3 Outlier Removal

As shown in Figure 2.1, the target points of a place with a continuous extent usually form an obvious cluster at that place with a few outliers surrounding it. There are several characteristics of the outliers that can help us remove them: 1) they lie remotely from the major cluster of target points; 2) they are thinly scattered around the major cluster; and 3) their total count is relatively small. So our goal is to find the major cluster and separate those thinly scattered points that are distant from it.

The Delaunay Triangulation Clustering (DTC) can meet our goal of identifying the major cluster in the target points. According to Eldershaw and Hegland (1997), the basic idea is to first construct neighbors for each target point based on Delaunay Triangulation and use edges (black solid lines in Figure 2.2(a)) to connect neighboring target points. For each pair of neighboring target points, if they are not close (i.e., their distance is beyond certain threshold, namely *cut-off distance c*),

they should not be in the same cluster, and their connecting edge is removed. Finally, sets of target points that are connected by remained edges (black solid lines in Figure 2.2(b)) form clusters. To choose a reasonable cut-off distance *c* for breaking edges, we assume the following two rules and the search procedure in Figure 2.2(c).

Rule (a): *c* must be no less than the sparsest target point's nearest neighbor distance *d* in Equation (2.2). This is because if *c* is less than *d*, the sparsest target point that is located in the most unpopular location will definitely be disconnected with any of its neighbors in target points, which can result in its isolation. This rule takes the location unpopularity issue into account when deciding the cut-off distance.

Rule (b): the largest cluster, denoted as the *major cluster*, in the clustering result under the cut-off distance *c* should contain more than 95% of the target points. This is because the distribution pattern of target points shows that the majority of target points form a cluster at the target place with only a few outliers surrounding it. The major cluster should consist of the majority of target points, and here we assume the number of them to be more than 95% of the target points. Empirically, this assumption works well across all eight places in this study.

When the major cluster is found, there are several isolated points and small clusters (e.g., A and B in Figure 2.2 (b)) that are quite near the major cluster. Since they do not have the outlier characteristic of being distant from the major cluster, they are not treated as outliers. A convex hull of the major cluster is created to capture these isolated points and small clusters. Finally, all other points or small clusters that do not intersect with the convex hull are removed from the set of target points. The remaining target points after the outlier removal step are denoted as the *cleaned points*.

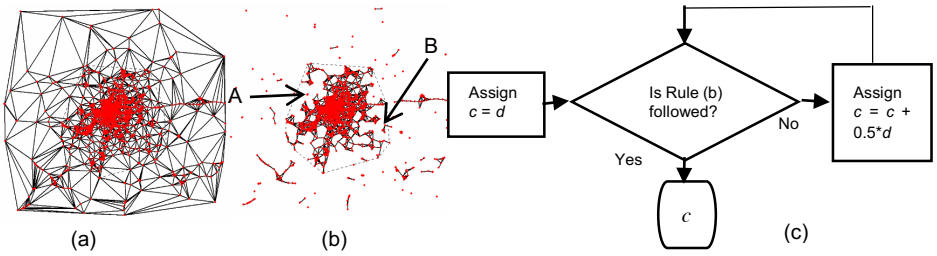


Figure 2.2 DTC with the target points (red dots) of Nashville: (a) edges connecting neighbors constructed by Delaunay Triangulation; (b) resulting major cluster with its convex hull (in dash line). (c) Flow chart of the search procedure for cut-off distance c.

2.3.4 Representativeness-Weighted KDE (RW-KDE)

We take the density surface approach to generate a raster surface representing vague place extents. The cleaned points provide limited observations about the location of a place. To measure the degree of other locations' inclusion in a place, interpolations can be made from the limited observations using the KDE. At each location to be estimated, the kernel estimator sums up the kernels centered at the cleaned points whose contributions decrease as distances increase (Silverman 1986). The traditional KDE method considers all observations with equal importance. We incorporate photo point representativeness into the KDE defined by Silverman (1986) using the following kernel estimator:

$$\hat{f}(X) = \frac{1}{h^2 \sum_{i=1}^{n} r_i} \sum_{i=1}^{n} r_i * K(\frac{X - X_i}{h}) \quad (2.3)$$

where n is the count of cleaned points; X_i is the coordinates of a cleaned point; X is the coordinates of a raster cell whose inclusion needs to be estimated; h is the smoothing bandwidth; r_i is the representativeness of cleaned point i calculated by Equation (2.1); and K is the quadratic kernel function cited from Silverman (1986) (p.76):

$$K(X) = \begin{cases} 3\pi^{-1} (1 - X^T \mathrm{X})^2 & if\ X^T \mathrm{X} < 1 \\ 0 & otherwise \end{cases} \quad (2.4)$$

For each set of cleaned points of a place, both KDE and RW-KDE are implemented using the same parameters. This ensures that their results are comparable. In order to both maintain a good resolution with details on the final raster surface and minimize the computational cost associated with high resolution, we choose about 1/300 of the width of a place study area to be the cell size of output surface. It is well-known that the smoothing bandwidth could have a significant impact on the output. For each place, we repeated KDE and RW-KDE using different bandwidths to evaluate the sensitivity of the improvements by RW-KDE method. We tested eleven bandwidths that are about 1/8, 3/16, 1/4, 5/16, 3/8, 7/16, 1/2, 9/16, 5/8, 11/16 and 3/4 of the width of the minimum bounding rectangle of the cleaned points. The reason for choosing this range is that almost all surfaces tend to be over smoothed when bandwidth approaches 3/4, and that for most places in this study, bandwidths smaller than 1/8 are too small for many locations within a place to find any cleaned point in its neighborhood when calculating kernel density estimate. Since the improvements can be observed across the eleven bandwidths, here we only present results based on three bandwidths: 1/4, 3/8, and 1/2.

2.4 Results

In this section, we use the official boundary and the surveyed boundary as references to compare the results derived from the RW-KDE method vs. the KDE method. All official boundaries come from the 2015 TIGER/Line® Shapefiles

provided by the U.S.Census Bureau (https://www.census.gov/cgi-bin/geo/shapefiles/index.php) and the park maps on the website of U.S. National Park Service (https://www.nps.gov/). Chinatowns are usually recognized and known for its distinct Chinese characteristics in building styles, signs, decorations, and pedestrians. These characteristics change notably between adjacent streets that are respectively inside and outside Chinatown. Thus, it is reasonable to manually draw a relatively objective boundary separating Chinatown from the surrounding areas, based on the Street View images from Google Maps.

All eight places were estimated by the steps described in Section 2.3: data download and preprocessing, quantification of representativeness of target points, outlier removal, and estimation of vague spatial extent using KDE vs. RW-KDE methods. Figure 2.3 shows the surfaces of vague extent generated by KDE vs. RW-KDE methods with the three selected bandwidths. All estimated values on the density surfaces are normalized to [0, 1] and linearly stretched for grey shades of [0, 255]. In the introduction section, we argue that location popularity could vary greatly across different areas of a place. To better know how strong the variation is, we take California as an example to estimate three popularity density surfaces (Figure 2.4) based on all photo points in the study area using the KDE, each of which corresponds to one of the three estimated KDE surfaces and RW-KDE surfaces of vague extent in Figure 2.3 under the same estimating parameters.

The popularity density surfaces show the advantage of San Francisco and Los Angeles areas which are two of the largest cities in the U.S. These surfaces are very similar to the estimated KDE surfaces of vague spatial extent in Figure 2.3. To know if a location's membership of California estimated from the traditional KDE method is correlated to the popularity density of that location, we plot these two variables for each location within the official boundary and calculate a Pearson's r correlation coefficient. The results indicate a very strong correlation which supports our early argument that location popularity may impact the estimates by the traditional KDE method. Take the popularity density surface with the bandwidth of 755,000 feet (Figure 2.4(c)) as an example, more than 65% of the state area have a popularity density below 0.2 in a scale of [0,1], and only 4% have a popularity density above 0.8. Among the low popularity-density locations (below 0.2), all of them have a membership value below 0.263 on the traditional KDE surface; 75% have a value under 0.13; 50% have a value under 0.075; and 25% have a value under 0.033. This means that the majority of locations within the official boundary are much unpopular than San Francisco and Los Angeles areas, and they are estimated by the traditional KDE approach to have a much lower possibility of being in California than the San Francisco and Los Angeles areas which have an estimate above 0.42. This is a significant deviation from the ground truth. A qualitative observation of the large dark area inside the official boundary tells the same story.

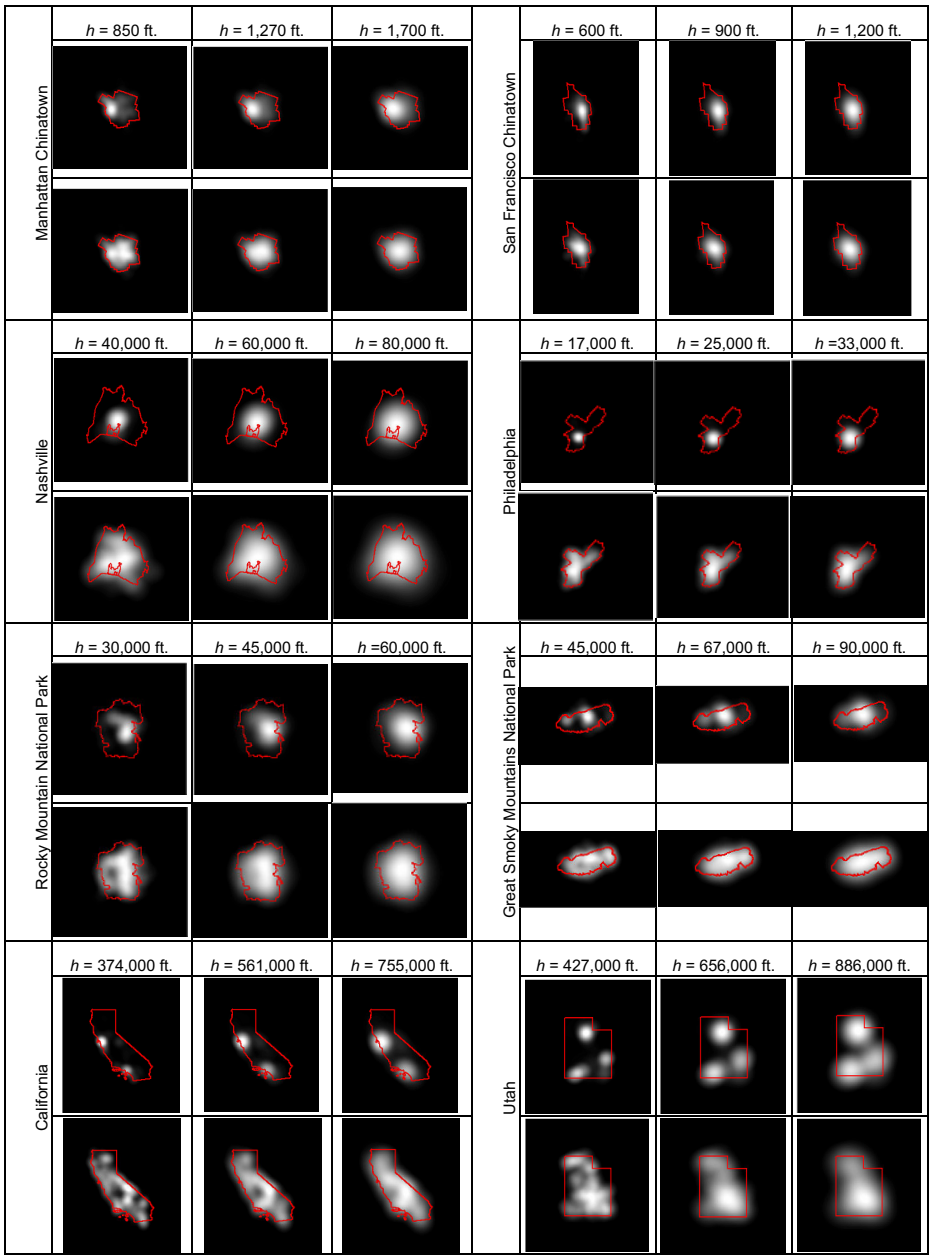


Figure 2.3 Vague spatial extents represented by KDE (top row of each place) vs. RW-KDE surfaces (bottom row) under different bandwidths (h). Reference boundaries are in red line.

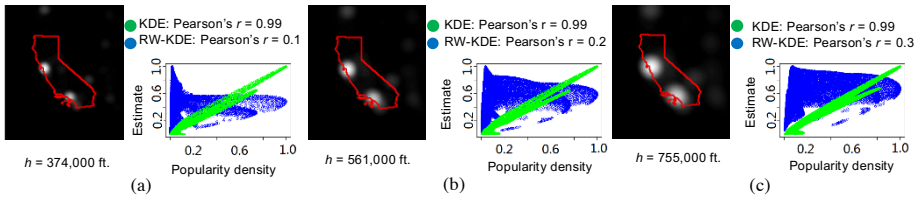


Figure 2.4 In each pair of (a) to (c), (left) a popularity density surface based on all Flickr photos using KDE; (right) a scatter plot of the estimates from the resulting KDE and RW-KDE surfaces against the popularity densities within the official boundary of California.

In contrast, by incorporating photo representativeness, the estimated values of RW-KDE surface are no longer correlated to a location's popularity density (see blue dots in Figure 2.4). Also among the same low popularity-density locations (below 0.2) within the official boundary, 75% of them have an estimated value above 0.4 on the RW- KDE surface; 50% have a value above 0.56; and 25% have a value above 0.73. The RW- KDE has greatly increased the estimated values at unpopular locations, better representing their membership of California. These improvements support our argument that we need to treat each photo tagged with a place name differently according to its location popularity; otherwise, less popular locations that obviously belong to a vague place would be underestimated in the derived geographic extent. Similar issues and improvements can be found in downtown vs. other areas of Philadelphia and Nashville, low vs. high accessibility areas in national parks, and across different bandwidths. Although there exists linear distributions of photo points in national parks, RW-KDE can take advantages of the limited number of sparse points that are away from the roads by assigning them a higher weight. It better represents the less popular areas that are often underestimated by the KDE method.

Moreover, based on qualitative observations of Figure 2.3, in the cases of Chinatowns, national parks, and Nashville, RW-KDE surfaces tend to produce the highest estimates near the center of reference boundaries. This is consistent with the common sense that the center of a place has the highest probability of membership, and that probability decreases when approaching the boundary line. However, the cores of traditional KDE surfaces tend to be distorted toward the most popular locations.

Both the KDE and the RW-KDE surfaces can produce a crisp boundary using a contour line with a threshold. To further quantify their performance, namely closeness to the reference boundary, we choose the commonly used measures of accuracy, recall, and precision. As defined by Schockaert et al. (2011), recall measures how much of the reference extent A can be covered by the estimated extent A', namely area(A∩A')/area(A); precision measures how much of the estimated extent correctly falls in the reference boundary, namely

area(A∩A')/area(A'); and accuracy measures the overall performance of the estimates, namely area(A∩A')/area(A∪A').

We use the rank of raster cells based on their density values in descending order rather than the value itself to derive the boundaries from the KDE surface and the RW-KDE surface, respectively. This is because the distributions of normalized density values from the KDE surface and the RW-KDE surface are quite different. If a threshold of density value is used to derive boundaries from these two surfaces, the returned boundaries could be very different in size. Since the size of derived boundaries can influence their comparisons with the reference boundary, we need to keep the two estimated boundaries comparable in size when examining their performance based on comparisons with the reference boundary. If a rank threshold β is used to derive the boundaries, the returned two sets of raster cells from the KDE surface and the RW-KDE surface will form two crisp boundaries with equal size but different shapes.

We calculate the recall, precision and accuracy measures for the boundaries generated with various rank thresholds from the KDE and the RW-KDE surfaces under the selected three bandwidths and plot them in Figure 2.5. These plots also include additional KDE surfaces estimated from the target points with outliers to see if it is the outlier removal step that contributes to the improvements. In Figure 2.5, a blue line (i.e., RW-KDE surface) is frequently above a line of the same type and other colors. That is, given the same threshold and bandwidth, boundaries derived from the RW-KDE method outperform those derived from the KDE method. RW-KDE method produces a less distorted representation of vague place boundaries than KDE approach. The outlier removal does not improve the performance of traditional KDE method, for red lines and green lines are almost overlapping with each other.

In the precision plots, many low precisions are found in crisp boundaries that are derived at a small rank threshold from KDE surfaces (see the left side of the precision charts of Philadelphia, California, Nashville in Figure 2.5). This means that many locations that are estimated by KDE to be most likely included in the target place are actually outside reference boundaries. This is consistent with the observed distortions in the derived KDE surfaces of Philadelphia, California, and Nashville in Figure 2.3. In the case of California, the overestimated areas that are outside the reference boundary have a high popularity density shown in Figure 2.4. This supports our point that the popular locations that are less likely to be located within the place could be overestimated and distort the boundary. In contrast, much higher precisions (close to 1) are found in crisp boundaries that are derived at the same small rank threshold from RW-KDE surfaces. That is, almost all locations that are estimated by the RW-KDE method to be most likely included in the target place are truly inside the reference boundaries.

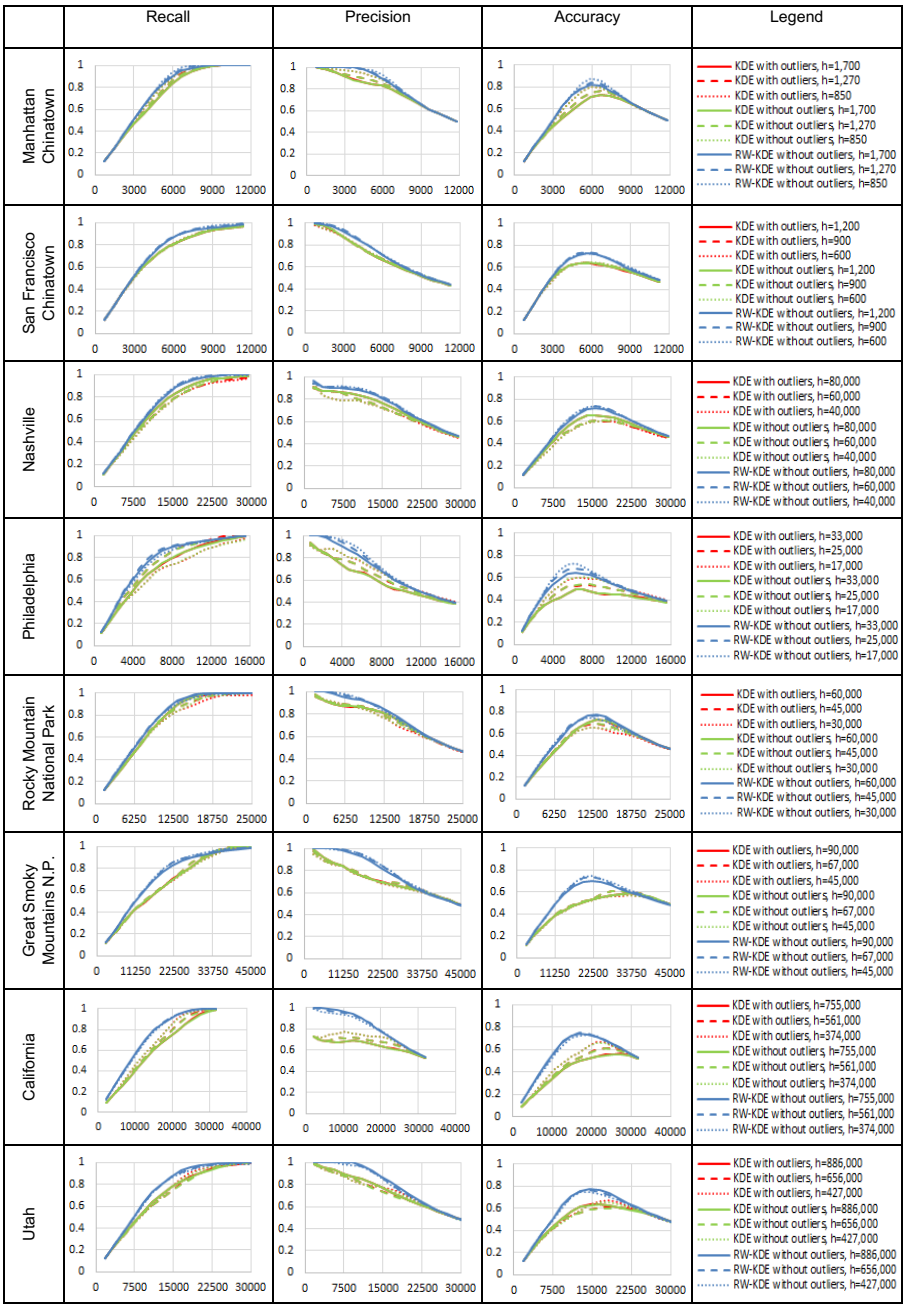


Figure 2.5 Recall, precision, and accuracy of the boundaries derived from the KDE and RW-KDE surfaces. The x axis represents the rank threshold β used to derive crisp boundary. The y axis in the second, third, and fourth columns represent the recall, precision, and accuracy measures, respectively.

2.5 Conclusions

Estimation of vague spatial extent of place names is important to GIS capability of handling places. Geotagged photos have brought great opportunities to estimate vague place extents. However, the challenge that photos are not randomly distributed makes it questionable if Flickr photos can be used to derive a good representation of vague place extent, not just to identify a single crisp boundary for a vague place.

Our analysis of the California example shows that, without a consideration of point representativeness, locations in less popular areas are likely to be underestimated using the traditional KDE method, and popular areas are likely to have overestimation that could distort the shape of derived spatial extents. With this challenge, this paper proposes a solution of assigning photo point representativeness based on their location popularity to improve the representation of vague place extents. Compared to the results derived from the traditional KDE method, the proposed RW-KDE method outperforms the traditional KDE, which is not subject to the kernel bandwidth change within a reasonable range. The locations in less popular areas that obviously belong to a target place are better estimated by the RW-KDE method to be comparable with those popular locations within the same target place. The locations in highly popular areas that do not belong to the target place are less likely to be estimated by the RW-KDE method to be part of the target place. The RW-KDE method also derives crisp boundaries with higher recall, precision, and accuracy measures, which quantitatively suggest a less distorted representation of place boundaries.

The major contribution of this paper is two folds: First, it addresses and proposes a solution for the aforementioned important challenge that has been widely recognized in the literature but not fully explored. The proposed method has been tested with eight places and produced better representation of vague place extents. Second, the improvements show that it is feasible to use geotagged Flickr photos to construct a good representation of vague place extents where a higher estimate indicates a higher probability of belonging to the target place.

Good matches between the estimated vague extents and the reference boundaries indicate that Flickr users' perception of the eight place extents is close to the reference boundaries. However, as suggested by (Cunha and Martins 2014), there could be some places whose derived vague extents are different from their official boundaries. More places will be examined in the future to find out useful patterns about peoples' perception of place, such as in what situations a place's derived extent is quite different from the official boundary.

The method proposed in this paper may not be suitable for places having multiple parts. This is because the outlier removal process assumes that the photos tagged with a target place name usually form one major cluster at that place. For places

containing disjoint parts or places sharing the same place name, additional considerations are needed in future research to detect the disjoint parts of a place. Then the proposed method in this study can be applied to generate a vague extent for each individual part.

References

Agnew, J.A., 2011. Space and place. In: J.A. Agnew and D.N. Livingstone, eds. The SAGE handbook of geographical knowledge. Thousand Oaks, CA: SAGE, 316–330.

Alani, H., Jones, C.B., and Tudhope, D., 2001. Voronoi-based region approximation for geographical information retrieval with gazetteers. *International Journal of Geographical Information Science*, 15(4), 287-306.

Arampatzis, A., et al., 2006. Web-based delineation of imprecise regions. *Computers, Environment and Urban Systems*, 30(4), 436-459.

Cohn, A. and Gotts, N., 1996. The 'Egg-Yolk' representation of regions with indeterminate boundaries. *In*: P. Burrough and A. Frank, eds. *Geographic objects with indeterminate boundaries*. Bristol, PA: Taylor and Francis, 171–187.

Cunha, E. and Martins, B., 2014. Using one-class classifiers and multiple kernel learning for defining imprecise geographic regions. *International Journal of Geographical Information Science*, 28(11), 2220-2241.

Dilo, A., De By, R.A., and Stein, A., 2007. A system of types and operators for handling vague spatial objects. *International Journal of Geographical Information Science*, 21(4), 397–426.

Downs, J.A. and Horner, M.W., 2012. Analysing infrequently sampled animal tracking data by incorporating generalized movement trajectories with kernel density estimation. *Computers, Environment and Urban Systems*, 36(4), 302-310.

Egenhofer, M.J. and Mark, D.M., 1995. Naïve geography. *In*: A. U. Frank, W. Kuhn, eds. *Spatial information theory a theoretical basis for GIS. Lecture notes in computer science vol. 988*. Berlin: Springer, 1–15.

Eldershaw, C. and Hegland, M., 1997. Cluster analysis using triangulation. *In*: *Computational techniques and applications, CTAC 97*. Singapore: World Scientific, 201-208.

Gao, S., et al., 2014. Constructing gazetteers from volunteered big geo-data based on Hadoop. *Computers, Environment and Urban Systems*.

Goodchild, M.F., et al., 1998. Fuzzy spatial queries in digital spatial data libraries. In: *Proceedings of the IEEE world congress on computational intelligence*. Anchorage: IEEE, 205-210.

Goodchild, M.F. and Hill, L.L., 2008. Introduction to digital gazetteer research. *International Journal of Geographical Information Science*, 22(10), 1039–1044.

Grothe, C. and Schaab, J., 2009. Automated footprint generation from geotags with kernel density estimation and support vector machines. *Spatial Cognition & Computation*, 9(3), 195-211.

Hollenstein, L. and Purves, R., 2010. Exploring place through user-generated content: using Flickr tags to describe city cores. *Journal of Spatial Information Science*, 1, 21-48.

Jones, C.B., et al., 2008. Modelling vague places with knowledge from the web. *International Journal of Geographical Information Science*, 22(10), 1045–1065.

Keßler, C., Janowicz, K., and Bishr, M., 2009a. An agenda for the next generation gazetteer: geographic information contribution and retrieval. *In*: *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems*. New York: ACM, 91–100.

Keßler, C., et al., 2009b. Bottom-up gazetteers: learning from the implicit semantics of geotags. *In*: K. Janowicz, M. Raubal, S. Levashkin, eds. *GeoSpatial semantics. Lecture notes in computer science vol. 5892*. Berlin: Springer, 83-102.

Li, L. and Goodchild, M.F., 2012. Constructing places from spatial footprints. *In*: *Proceedings of the 1st ACM SIGSPATIAL international workshop on crowdsourced and volunteered geographic information*. New York: ACM, 15–21.

Liu, Y., 2014. *A study of colloquial place names through geotagged social media data*. Thesis (master), University of Tennessee.

Martins, B., 2011. Delimiting imprecise regions with georeferenced photos and land coverage data. *In*: K. Tanaka, P. Fröhlich, K. Kim, eds. *Web and wireless geographical information systems. Lecture notes in computer science vol. 6574*. Berlin: Springer, 219-229.

Montello, D.R., et al., 2003. Where's downtown?: Behavioral methods for determining referents of vague spatial queries. *Spatial Cognition & Computation*, 3(2-3), 185–204.

Montello, D.R., Friedman, A., and Phillips, D. W., 2014. Vague cognitive regions in geography and geographic information science. *International Journal of Geographical Information Science*, 28(9), 1802-1820.

Parker, J.K. and Downs, J.A., 2013. Footprint generation using fuzzy-neighborhood clustering. *Geoinformatica*, 17(2), 285-299.

Purves, R., Clough, P., and Joho, H., 2005. Identifying imprecise regions for geographic information retrieval using the web. *In*: *Proceedings of the GIS research UK 13th annual conference*. Glasgow: University of Glasgow, 313-318.

Relph E., 1976. *Place and placelessness*. London: Pion.

Tuan, Y.-F., 1977. Space and place: the perspective of experience. Minneapolis: University of Minnesota Press.

Schockaert, S., Smart, P.D., and Twaroch, F.A., 2011. Generating approximate region boundaries from heterogeneous spatial information: an evolutionary approach. *Information Sciences*, 181(2), 257-283.

Schockaert, S., De Cock, M., and Kerre, E.E., 2005. Automatic acquisition of fuzzy footprints. *In*: R. Meersman, Z. Tari, and P. Tari, eds. *On the move to meaningful internet systems 2005: OTM 2005 workshops. Lecture notes in computer science vol. 3762*. Berlin: Springer, 1077-1086.

Silverman, B.W., 1986. Density estimation for statistics and data analysis. London: Chapman and Hall.

Twaroch, F.A., Jones, C.B., and Abdelmoty, A.I., 2009. Acquisition of vernacular place names from web sources. *In*: I. King and R. Baeza-Yates, eds. *Weaving services and people on the world wide web*. Berlin: Springer, 195-214.

Chapter 3

Estimating the Spatial Extent of a Place with Disjoint Regions Using Flickr Photos

Abstract

Flickr photos' uneven spatial distribution is a major challenge when using such data to estimate the spatial extent of places that have no formally defined boundary. There is limited research that considers this issue when approximating the vague extent of a place with disjoint regions. A place with disjoint regions is defined as a place whose spatial coverage is not a single continuous surface but consists of multiple disjoint extents located in different geographical areas. To fill the research gap, this paper proposes a method named *representativeness-weighted kernel density estimation for disjoint extents* (RW-KDE-FDE). This method is tested with three place names in study areas of different scales. It successfully determines the places' disjoint extents and improves their spatial representation compared to the results from the traditional kernel density estimation (KDE) approach.

3.1 Introduction

Humans perceive, reason and convey the geographic world differently from the ways current Geographic Information Systems (GIS) do; the former usually take a platial perspective while the latter generally take a spatial perspective (Goodchild 2015). A well-known example is that place names, rather than abstract coordinates as seen in GIS, are frequently used in daily life to specify spatial features. This difference impedes ordinary people's use of powerful GIS functions and GIS services' inclusion of human geographic knowledge (Egenhofe and Mark 1995). Representing the spatial extent of places, especially those without a formally defined boundary, is important when converting between space and place for the development of human-centric GIS.

Representation of places on a map or in a GIS database is more complex and challenging than placing a dot or a pushpin on a map, especially when the place is polygonal and its extent is unknown and vague. Great efforts have been made to model vague place extents in GIS (e.g., Dilo et al. 2007, Guo et al. 2008) and to derive them based on different data sources (e.g., Arampatzis et al. 2006, Montello et al. 2003). Recent literature shows increasing interest in using Flickr geotagged photos to determine the vague extent of places (e.g., Li and Goodchild 2012, Cunha and Martins 2014) because of Flickr data source's advantages such as high volume, low cost, and a long time span. However, an important challenge is that Flicker photos' uneven spatial distribution could have a great impact on the derived extents. Taking the commonly used kernel density estimation (KDE) as an example, a higher density of photos tagged with a target place name does not necessarily mean a higher probability of being that place, but could simply correspond with the underlying high availability of photos.

This challenge has been recognized in several studies (e.g., Grothe and Schaab 2009, Hollenstein and Purves 2010) but has not been sufficiently addressed. Chen and Shaw (2016) tried to deal with this challenge by modeling Flickr photo's

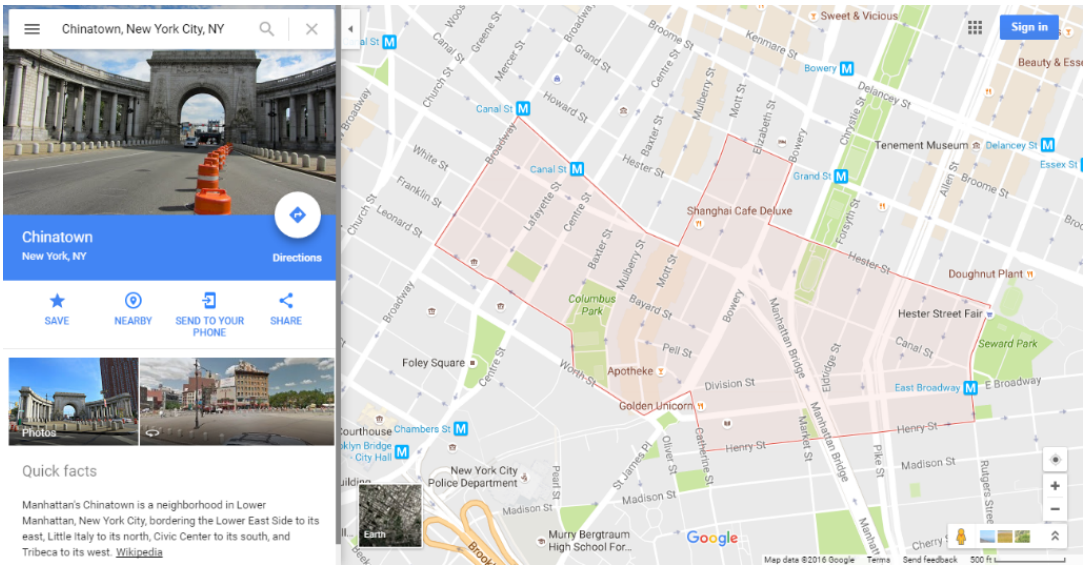
representativeness based on location popularity and proposed a representativeness-weighted KDE (RW-KDE) method, which produced less distorted place extents. However, this method is unsuitable for a place with disjoint extents because its outlier removal process and parameter selecting procedure assume that only one continuous extent corresponds to a place name in a study area. However, in the real world, the same place name could be used to specify different regions that are spatially disjoint as the study area's scale becomes large enough (Feick and Robertson 2015a). The scale of study unit is important when harvesting place-related information from Flickr data (Feick and Robertson 2015b). For example, one region is named Chinatown when the study area is at the scale of New York City's Manhattan borough, while at least three disjoint regions are named Chinatown (located in the boroughs of Manhattan, Brooklyn and Queens) at the scale of New York City. Since places can change over time given a study area, the number of regions sharing the same place name is usually dynamic and not well known. Figure 3.1 lists different results returned from the query of Chinatown, New York City. Wikipedia delivered better information about all of New York City's neighborhoods called Chinatown than Google Maps, which identified the one in Manhattan. These different results indicate the importance and challenge of determining a place name's disjoint extents. Thus, this paper aims to identify an effective approach to estimating a place's disjoint vague extents while considering Flickr photos' uneven spatial distribution.

We present a method named *representativeness-weighted kernel density estimation for disjoint extents* (RW-KDE-FDE), which first determines the geographical areas into which disjoint extents may fall based on a spatial scan statistic approach (Kulldorff 1997) and then separately estimates each extent in the local area using an improved RW-KDE approach (Chen and Shaw 2016). This method was tested with three place names in study areas of different scales. The results show successful determination of disjoint extents and improved representation of place extents.

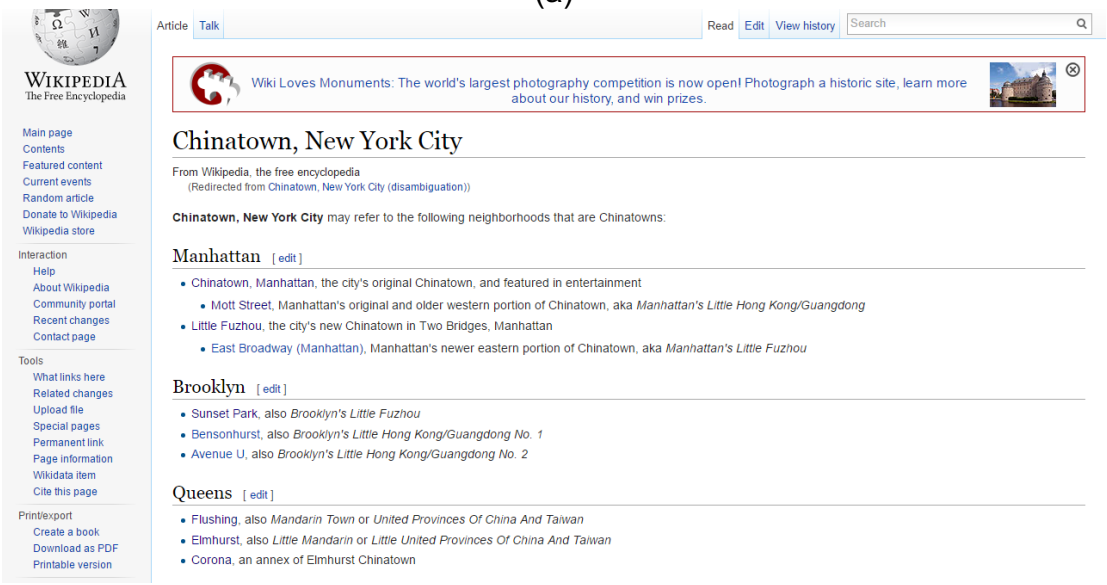
The remainder of this paper is organized as follows. Section 3.2 reviews related work on space, place and GIS; extraction of place-related information from volunteered geographic information (VGI); and vernacular place names' spatial extent representation. Section 3.3 discusses the RW-KDE-FDE method, followed by the results of testing with the three selected place names in Section 3.4. In Section 3.5, this paper is concluded with contributions, limitations and future research directions.

3.2 Related Work

Although the definition of *place* is still vague after a long history of discussions in human geography and related fields (Goodchild 2011), a place is frequently seen as a product of endowing space with human experiences, perceptions, values, and



(a)



(b)

Figure 3.1 Search results of Chinatown, New York City from (a) Google Maps and (b) Wikipedia.

social meanings through collective human-environment interactions (Tuan 1977, Casey 1997, Agnew 2011). Because of the inherent vagueness of place, the concept of place benefits little from today's digital computers with high accuracy and efficiency in terms of calculation, visualization, analysis and distribution (Goodchild 2011). Much effort has gone into exploring place in the context of GIS.

For practice purposes, Agnew (2011) modeled places to be consisting of three dimensions: location, locale, and sense of place. The location dimension relates to the counterpart of places in space, such as geolocation and spatial extent, which this paper explores. The locale dimension refers to the social properties of place, such as homes and workplaces. The sense-of-space dimension deals with individuals' feelings for a place. Most research aiming to bring place into GIS involves the location dimension and gradually moves into the locale dimension (e.g., Purves and Derungs 2015, McKenzie et al. 2015). An example of addressing the location dimension is the application and development of online worldwide digital gazetteers (e.g., GEOnet Names Server, Alexandria Digital Library, and Getty Thesaurus of Geographic Names), linking place names to their geographic locations that can be represented in GIS. However, the lack of large numbers of vernacular place names and spatial extents is a major deficiency in current digital gazetteers. Thus, acquiring place-related information is critical to enriching gazetteers and addressing all the dimensions of place.

The past decade's "convergence of GIS and social media" (Sui and Goodchild 2011, p.1737) has created tremendous volumes of dynamic volunteered geographic data, which provide an unprecedented opportunity to obtain place-related information (Goodchild et al. 2016). Related work includes the following: distinguishing place semantics from non-place semantics in texts (e.g., Rattenbury and Naaman 2009, Mackaness and Chaudhry 2013); inferring the location of a place that a set of words describes (e.g., O'Hare and Murdock 2013); extracting a meaningful urban area of interest (e.g., Hu et al. 2015); and inferring the place type of urban points of interest (e.g., McKenzie et al. 2015). Among the many aspects of place-related information that can be derived from VGI, the spatial extent of places has gained much attention.

Before VGI was used, many data sources were explored to delineate the vague extent of places. For example, Montello et al. (2003) collected 36 individuals' drawings of Downtown Santa Barbara's boundary with different levels of confidence. They then aggregated the drawings to generate a frequentist representation of the Downtown area. Considering survey data's advantages, such as quality control and sampling strategy, some recent work still uses this kind of data to explore places' extent and the patterns of people's cognitive regions (e.g., Lüscher and Weibel 2013, Montello et al. 2014). However, data collection's high cost makes using survey data impractical for a large-scale exploration of human perception of places in a timely manner. For easier acquisition of place-related

data, Schockaert et al. (2005), Arampatzis et al. (2006) and Jones et al. (2008) turned to web-sourced documents for references (e.g., stores, hotels and counties) that were described to be inside, outside, containing the place under study, or having other relations to that place. Based on the references' geolocations and their relationship to the target place, the spatial extent was estimated and represented in the form of density surface, fuzzy set, or crisp boundary. However, the process of selecting valid references and determining their geolocations is usually challenging and tends to produce errors (Martins 2011).

Using geotagged photos to extract the spatial extent of places is becoming a trend in this VGI era. From a geographic perspective, geotagging and textual tagging a photo is a process in which Flickr users explicitly document their experiences with and their understandings and feelings about the place where the photo is located. Thus, the large number of photo geolocations and associated text tags (containing place names) is a good direct-data source for harvesting human knowledge of the spatial extent of places. Grothe and Schaab (2009) and Cunha and Martins (2014) used the one-class support vector machine approach to find the globally maximum-density cluster of the photo points tagged with the target place name. The cluster's region was treated as the target place's spatial extent. The derived place boundaries were crisp and did not reflect the inherent vagueness of places. KDE has been widely used in the literature (e.g., Hollenstein and Purves 2010, Martins 2011, Li and Goodchild 2012) to represent the vague boundary of places, but has encountered the challenge of Flickr photos' uneven spatial distribution as mentioned in the introduction section. That is, the resulting KDE surface may correlate with the Flickr photos' overall distribution.

Chen and Shaw (2016) argued that each Flickr photo point's representativeness should be different because the location popularity (or photo availability) varies across the entire study area. Thus, they proposed a method to model each photo point's representativeness; this method improved the KDE-based representation of the vague boundary of places. An outlier removal process was included in their research to reduce the impact of photos tagged with the target place name but not actually located within the target place. However, this outlier removal process is incapable of dealing with places with two or more disjoint extents, thus creating difficulties for properly estimating their vague extents. Little research is available that both estimates the vague extent of a place with disjoint regions and considers the effect of Flickr photos' uneven distribution.

A target place name's photo points are predominantly located at that place and exhibit a clustering pattern in terms of density. Two or more clusters could exist if the target place has disjoint extents. Clustering techniques can be used to first determine where and how many disjoint extents exist. Then each extent can be approximated separately using existing estimation methods. The literature offers a variety of spatial clustering approaches, such as K-means (MacQueen 1967),

DBSCAN (Ester et al. 1996), and hierarchical clustering. Two challenges could arise in using these approaches to detect the disjoint extents. First, some parameters (e.g, K-means' number of clusters and DBSCAN's minimum number of points in neighborhood) need to be predefined and largely influence the results. Second, these approaches cannot take into account Flickr photos' underlying distribution; thus, may be unable to distinguish outlier clusters. As a specific type of VGI, Flickr data inevitably contain some outliers because of the lack of quality assurance (Goodchild and Li 2012). The outliers could grow into clusters at locations where the numbers of overall photos are extremely high. Thus, it is important to distinguish the outlier clusters from the significant clusters representing the place extents. Figure 3.2 presents an example in which the observed cluster in the red rectangle area is not actually the target place, Chinatown; instead, the cluster reflects the area of the hottest tourist attractions in New York City's Manhattan borough.

The spatial scan statistic approaches (e.g., Openshaw et al. 1987, Kulldorff 1997, Kulldorff et al. 2006) are better for addressing the above challenge. They use search windows moving across the entire study area to look for regions that have anomalously high rates of observations against a random process. Kulldorff (1997) presented a mature spatial scan statistic, which has been widely used in crime and health fields to identify crime hotspots and epidemic bursts of statistical significance. There is limited research applying this kind of method to VGI data that possess very different characteristics from traditional datasets such as census-based crime and disease reports. In this paper, using the spatial scan statistic to estimate the disjoint extents of places is also a means of exploring a traditional spatial statistic approach's applicability to new type of data in the VGI era.

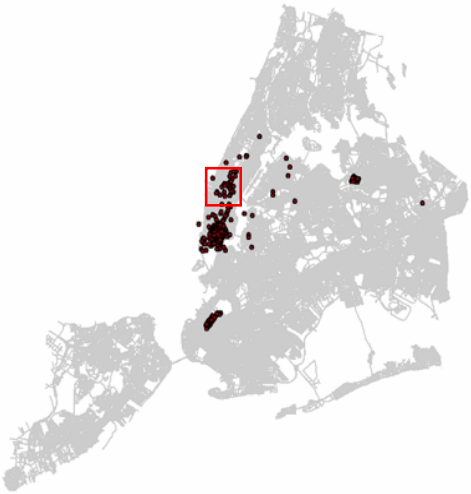


Figure 3.2 Spatial distribution of the Flickr photos tagged with "Chinatown" in New York City.

3.3 Methodology

3.3.1 Data Acquisition and Preprocessing

To evaluate RW-KDE-FDE's performance with different place types and study-area scales, we selected three place names as case studies and defined their study areas as follows: *Square Park* with an official boundary within Manhattan, New York City at the urban borough scale; *Chinatown*, which does not have an official boundary within New York City, at the urban city scale; and *National Park* with an official boundary within California at the state scale. We denote these places as *target places* in the remainder of this paper.

For each target place, its study area's envelope (i.e. Manhattan, New York City, or California) was used to download the metadata of all geotagged photos using Flickr search API, regardless of whether their textual tags contained the target place name. Only the photos within each study area's official boundary were used for estimations. For Square Park and National Park, the Flickr photos posted between January 2013 and December 2015 and between July 2013 and January 2015, respectively, were downloaded to test whether partial Flickr photos could be appropriate for extracting place extents. For Chinatown, we chose the time span between January 2013 and December 2015 to correspond with the time period during which the Google Street View images were collected in the regions covering Chinatown. These street view images served as references for us to manually draw Chinatown's boundary, which doesn't officially exist but which could be used as a benchmark reference to validate the place extent derived in this study. Keeping the temporal comparability between the derived extent and its reference boundary is important.

Flickr users may upload photos in bulk and tag them with identical geolocations and textual tags (Hollenstein and Purves 2010). To reduce the data bias this phenomenon introduces, the same user's photo redundancy at the same location should be eliminated. Within each group of Flickr photos sharing an identical geolocation and user, only one photo was randomly chosen and kept, while the rest were removed from the dataset. Since some bulk-uploaded photos were not exactly identical in geolocation but very close to each other (about 10 to 100 meters away), they may also introduce biases and should be regarded as redundancy when the study area is large. Thus, we treated the same user's photos as redundant when their geographic coordinates were identical in decimal degrees to a precision of four decimal places for Chinatown and Square Park, and a precision of three decimal places for National Park. The term *all points* refers to the geolocations of the set of photos without redundancy. Among the all points, those tagged with the target place name are denoted as *target points*. The proposed RW-KDE-FDE method consists of three major steps: detection of disjoint extents, outlier removal in the local study area, and individual extent's density surface estimation.

3.3.2 Detection of Disjoint Extents

The name of a place is usually more likely used on Flickr photos located in that place than on photos located elsewhere. Thus, we formalized the problem of estimating the disjoint extents' locations as detecting the photo bursts tagged with a target place name. The case studies described in this paper tested the validity of this assumption and formalization. As a widely-used burst-detection method, the spatial scan statistic (Kulldorff 1997) can consider the Flickr photos' underlying spatial distribution. Thus, this method was applied to the dataset of all points and target points to determine the target points' bursts that might reflect a place's disjoint extents.

Many overlapping circles of varying sizes were created and centered on each target point, representing target points' candidate clusters (See Figure 3.3). Compared to other window shapes such as ellipse, circular windows cost less in computation and produced good results in our case studies. In each candidate cluster (i.e., circular window), the following test statistic, namely the logarithm of the likelihood ratio (LLR) (Kulldorff 1997), can be calculated to measure the probability of current cluster's having an unusually high rate of target points that is not by chance:

$$LLR = \begin{cases} \ln\left(\frac{n_C}{\mu_C}\right)^{n_C} + \ln\left(\frac{n_T - n_C}{n_T - \mu_C}\right)^{n_T - n_C} & if\ n_C > \mu_C \\ 0 & otherwise \end{cases} \quad (3.1)$$

where n_C is the actual count of target points within the circle; μ_C is the expected occurrences of target points based on the underlying photo availability (i.e., the number of all points in the circle) under the assumption of random Poisson process; and n_T is the total number of target points in the study area. This test statistic is calculated over all the study area' circular windows. The one with the maximum LLR indicates the most likely cluster of photos tagged with the target place name, namely the most likely extent of the place.

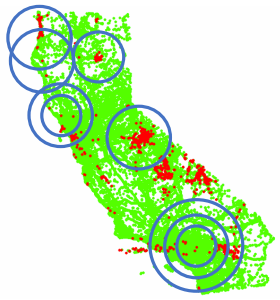


Figure 3.3 Examples of circular windows (blue circles) centered on the target points tagged with "national park" (red dots). The green dots represent the set of all points in California.

As suggested by Kulldorff (1997, 2015), a Monte Carlo hypothesis testing (Dwass 1957) was used to assess whether the most likely cluster is statistically significant. We generated 9999 Monte Carlo replications of our dataset under the assumption of random Poisson distribution. For each replicated dataset, the process described in the last paragraph was repeated to find a maximum LLR. Then the maximum LLR's rank derived from the real dataset was found to be r by comparing it with the 9999 maximum LLRs generated from the replicated datasets. A p-value was calculated as r divided by 10000 and compared to the significance level α. If the p-value of the most likely cluster found in the real dataset is equal to or less than α, the probability of observing such a cluster by chance is no more than α. In other words, the cluster is statistically significant at the α level; thus, it is considered to represent the extent of the target place. In many statistical applications, the typical values for α are 0.05 and 0.01. In this study, we tested three α values (0.05, 0.01 and 0.001) to assess α's influence on the results.

The significant most likely cluster identified above represents only one of the many disjoint extents of a target place. To identify other disjoint extents, we had to detect the clusters that also have high likelihoods but are secondary to the most likely cluster. We adjusted Kulldorff's (2015) iterative method, which repeats the complete process of identifying the most likely cluster based on the dataset from which the target points and all points of the clusters detected in previous iterations have been removed. The iterations stop when the last-found most likely cluster's p-value is greater than α. Among the clusters generated in iterations (except the one stopping the last iteration), those overlapping the ones derived in earlier iterations were removed because they are usually areas containing or joining the clusters that are more important and already identified. Including them would impede determining the number of real clusters. Then the remaining clusters' LLRs and p-values were recalculated based on the original dataset and the 9999 maximum LLRs generated in the first iteration. Only those clusters whose p-values are less than or equal to α are considered significant and may represent the target place's disjoint extents.

Note that the overlapping circles for detecting significant clusters should have an upper size limit. As Kulldorff (2015) suggested, a circular window should not contain more than 50% of the all points in the study area. The reason is that given the suitability of Equation (3.1), when a cluster is larger than 50%, its LLR does not indicate the unusually high rate of target points within the circular window, but instead indicates the unusually low rate of target points outside the circular window. Besides the 50% rule, a geographical maximum circle size s should be specified to avoid big clusters. The reason is that as the study area's scale increases, some real clusters become relatively closer to each other and tend to be identified as one big cluster in the detection process. However, when zoomed into the big cluster, the contained individual clusters become more observably separated and

identifiable. Thus, we assumed the search procedure diagrammed in Figure 3.4 to choose a reasonable s for each case study.

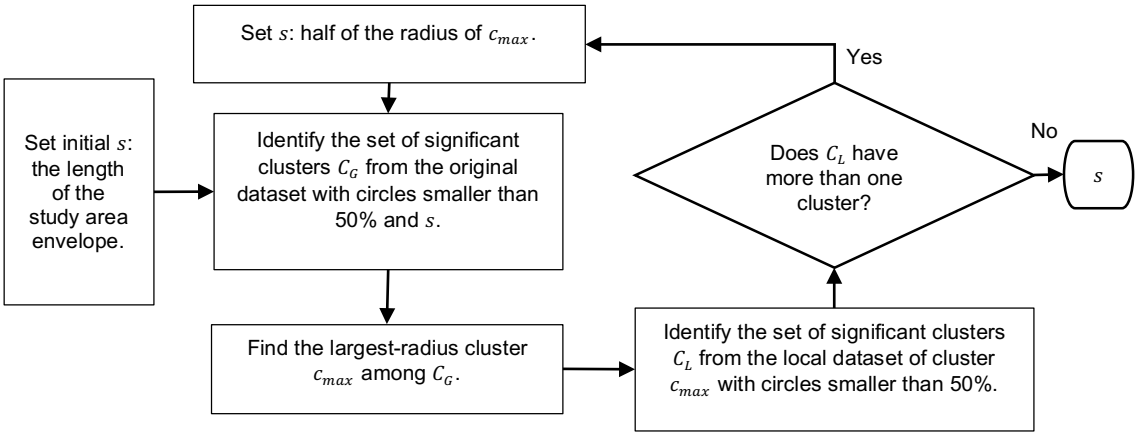


Figure 3.4 Search procedure for the maximum circle size.

In this subsection, the number of significant clusters found for each place name indicates the number of the target place's disjoint extents. The circular clusters are only rough estimates of the disjoint extents, but can be used to separate the global study area (i.e. Manhattan, New York City or California) into small local areas where each extent can be individually estimated.

3.3.3 Estimation of Individual Vague Extent in Local Study Area

The real shape of a place's extent is much more irregular than a simple circle. Each detected significant cluster and its surrounding area form the region that the actual extent may fall into and also include noises near the place. To approximate the disjoint extents of a target place, we first defined a local study area for each individual extent as a circular region centered on a detected significant cluster and whose radius is two times as large as the cluster's (see Figure 3.5). Then within each local study area, a modified RW-KDE method was applied to estimate the vague extent in that area.

RW-KDE's purpose is to first quantify the representativeness of each photo point to be inversely proportional to its location's popularity, which is measured by counting all points in that location. A location is a cell on a regular grid covering the study area. An outlier removal process is also included in RW-KDE to remove the photo points tagged with a target place name but not located in that place. Then the photo points' representativeness is incorporated into a quadratic kernel density estimation of the target points without outliers. Finally, the vague extent is represented by the resulting density surface where a higher value indicates a higher probability of being the target place. More details about the RW-KDE method can be found in Chen and Shaw's (2016) work and are not discussed in

this paper. The original outlier removal process was improved in this paper for better estimations.

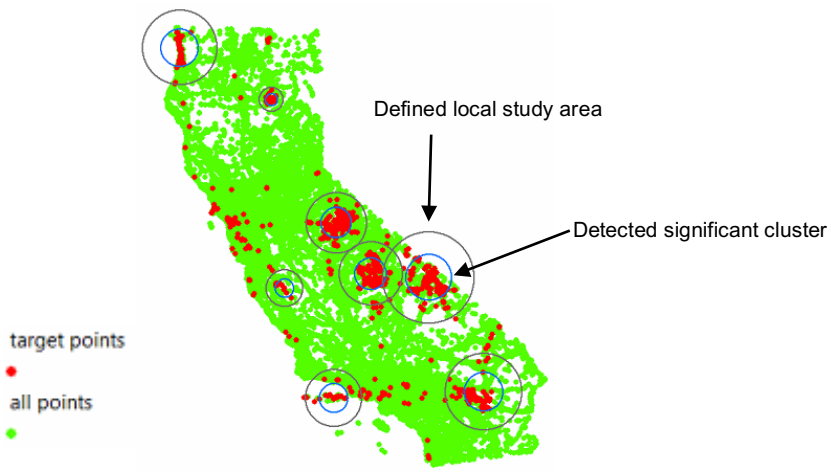


Figure 3.5 Example of the local study areas (black circles) centered on the detected clusters (blue circles) of target points in California.

The target points outside all local study areas do not form any significant cluster; thus, those points are regarded as outliers. Inside each local study area, Chen and Shaw's (2016) outlier removal process assumes that there should be a major cluster of target points representing the target place and surrounded by a few sparse and distant outlier points. The major cluster can be found based on Delaunay triangulation clustering (DTC) (Eldershaw and Hegland 1997). In each local study area containing an individual extent, a Delaunay triangulation can be used to construct the neighborships among target points. When two neighboring target points are not close enough (i.e., the length of the edge connecting them exceeds threshold distance c), their neighborship edge is removed, indicating they are not in the same cluster. Then the largest set of target points still connected by the remained edges can be found, and those target points falling within its convex hull are regarded as the *major cluster*. We modified the original outlier removal process by changing the rules of choosing a reasonable threshold distance c:

Rule (1): Considering varying location popularity, c should be at least equal to the nearest neighbor distance of the sparsest target point, which is defined by $d = max_{t \in T}\left(min_{a \in A-\{t\}} dist(t,a)\right)$ where T is the set of target points denoted by t and A is the set of all points denoted by a within the local study area. The sparsest target point is defined as the target point with the largest nearest-neighbor distance to all other photo points. This rule is the same as the definition in Chen and Shaw's (2016) work.

Rule (2): The major cluster's LLR under the threshold distance c should reach the highest point as c increases from d by replacing the circular window with the major cluster's convex hull. Note that μ_C and n_T in Equation (3.1) for calculating LLRs are determined based on the dataset in the global study area to meet the 50% criteria. The major cluster with the highest LLR indicates the most likely burst of photos tagged with the target place name in the local study area, thus assumed to better represent the extent. This rule replaces Chen and Shaw's (2016) rule that the major cluster should contain no less than 95% of the target points in the study area. Using an arbitrary value of 95% could be less robust than dynamically searching for the most likely cluster. The results of outlier removal based on these two rules are compared in Section 3.4.

Thus, the search procedure for threshold distance c was modified as shown in Figure 3.6. After outliers were removed, one additional step was taken to examine whether the major cluster is significant in terms of user contribution. Sometimes even a cluster is found to be statistically significant in Subsection 3.3.2; the cluster's evidence of representing the place extent becomes weak if the photos of that cluster were contributed by a very limited number of users. Thus, we assumed that if fewer than three unique users contributed a major cluster's photos, the cluster and its local study area were no longer considered significant and thus were excluded from representing place extent. Finally, each significant major cluster's target points were used to generate a density surface based on the RW-KDE method. Their representativeness and relevant parameters for producing the surface were determined in the local study area.

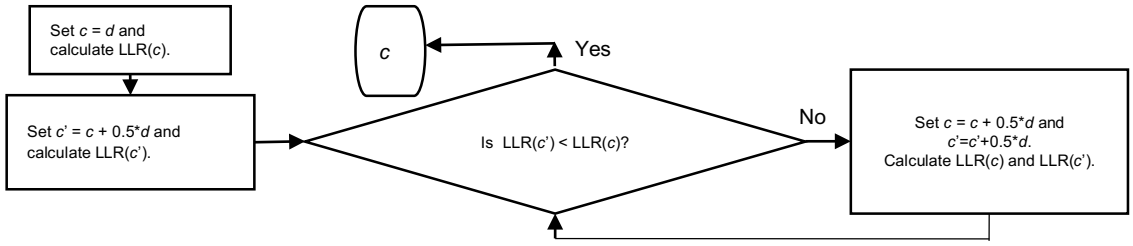


Figure 3.6 Search procedure for threshold distance *c*.

3.4 Results

For each place, we used its official boundary or surveyed boundary as the reference to evaluate RW-KDE-FDE's performance in determining the number of disjoint extents, removing outlier points and representing vague extent. According to the New York City Department of Parks & Recreation's website (https://www.nycgovparks.org/), five disjoint areas are named Square Park in Manhattan. The U.S. National Park Service's website (https://www.nps.gov) lists eight regions named National Park in California. Square Park's and National Park's official boundaries are from these two websites. We also referred to Wikipedia for

information on all Chinatown neighborhoods; and then by observing Google Street View images, we manually delineated these neighborhoods' boundaries enclosing regions with strong Chinese characteristics. The estimation of disjoint extents based on spatial scan statistic described in Section 3.3.2 was implemented using the free software SaTScan™ developed by Kulldorff and Information Management Services Inc. (2015). Using the steps described in Section 3.3, we analyzed all three place names.

3.4.1 Detected Significant Clusters for Disjoint Extents

The choice of significance level α is important when applying Monte Carlo hypothesis testing to determine whether a cluster of target points is statistically significant against the null hypothesis of random process. We tested three commonly used α values to see how they influence estimating disjoint extents. Table 3.1 presents the cluster detection's performance with different α by comparing the result with the ground truth. We used the number of national parks indicated on the official website as the benchmark reference in Table 3.1 although some extents have a tiny peripheral area separated from them (see the Lassen Volcanic National Park's reference boundary with its small disconnected area A in Figure 3.7(a)). Given our study area's large scale, we neither treated those tiny areas as independent nor counted them in the number of disjoint extents. Future work can focus on this scenario to determine small peripheral areas separated from a place's major extent. Although Wikipedia identifies as many as eight Chinatown neighborhoods, many of them are linearly located along streets (see red single lines in Figure 3.7(c)). Only three neighborhoods have a considerable region size and thus were estimated in our study.

Table 3.1 indicates good estimation results across different places and α values. All disjoint extents were successfully identified except one area named Square Park, which is the area B in Figure 3.7(b). Only two photos were tagged as Square Park in this area, which were not enough to be a cluster. However, the total number of all photos taken in this area is not low, possibly indicating that people do not tend to call this place by its official name. All of Chinatown's disjoint extents were successfully determined, although one extent was identified as two clusters (see area C in Figure 3.7(c)). The reason could be related to the big difference in shape between the circular window and the real extent that is long and narrow. Two small circles side-by-side can better exclude locations with low target-point rates than a single large circle. Given place extents' irregular shape, using an irregular scan window in future work might be more appropriate for better determining a place's real extent.

When a stricter significance level was used, fewer regions were wrongly estimated to be place extents (see the last column of Table 3.1 for National Park). In those

Table 3.1 Performance comparison of determining disjoint extents under different significance levels. (The number with * indicates that among the detected significant clusters, there are two clusters corresponding to one place extent.)

Place name	Significance level α	Number of detected significant clusters	Real number of regions with the place name	Errors	
				Number of real regions that are not detected	Number of detected clusters not corresponding to any extents
National Park	0.05	12*	8	0	3
	0.01	10		0	2
	0.001	9		0	1
Chinatown	0.05	4*	3	0	0
	0.01	4*		0	0
	0.001	4*		0	0
Square Park	0.05	5	5	1	1
	0.01	5		1	1
	0.001	5		1	1

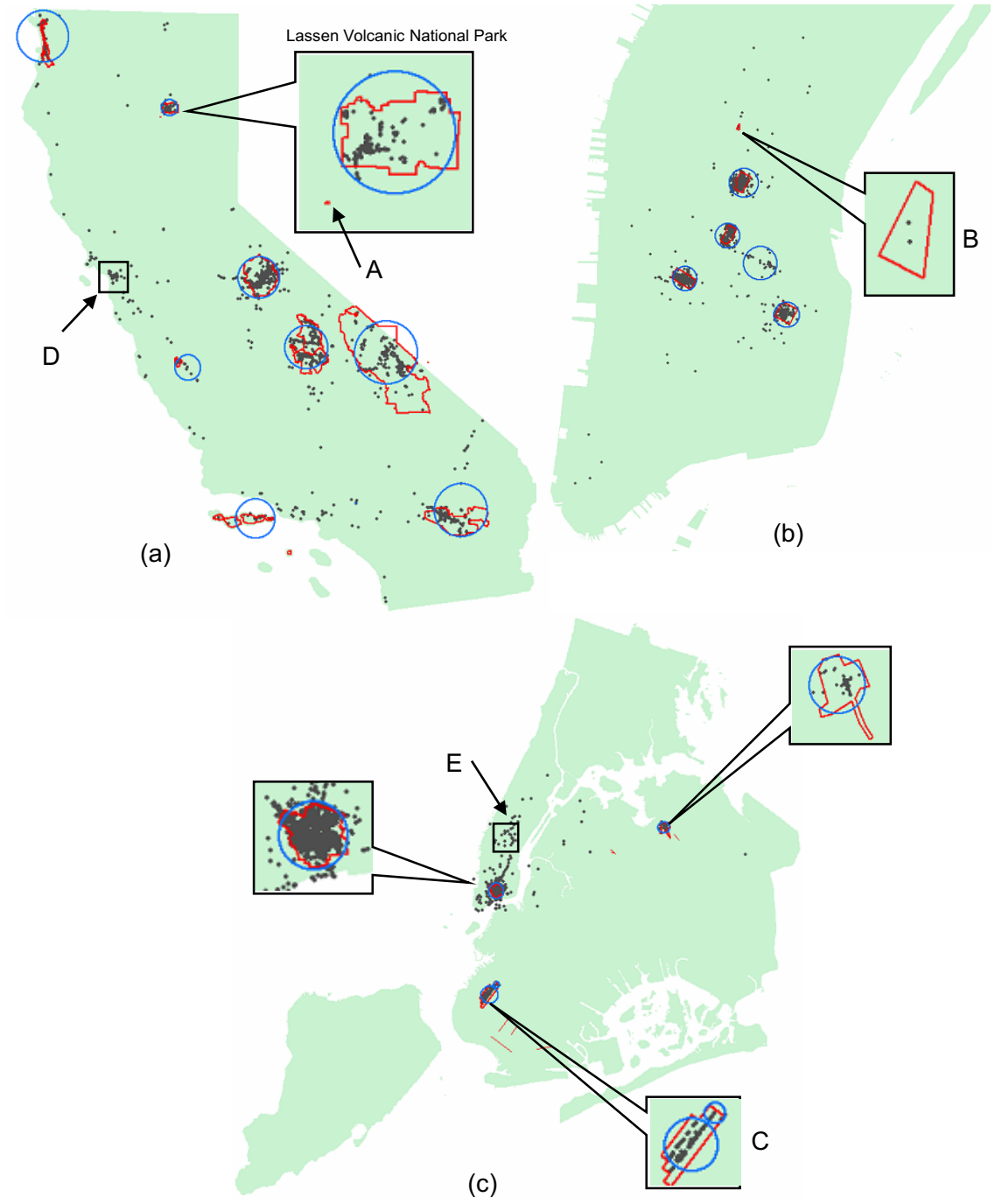


Figure 3.7 Detected significant clusters (blue circles) of disjoint extents at α=0.001 vs. reference boundaries of (a) National Park, (b) Square Park and (c) Chinatown. Grey dots are photo points with a target place name. Reference boundaries are in red lines, which applies to all following figures.

wrongly estimated regions, few people contributed photos; but one person uploaded numerous photos tagged with the incorrect target place name. This error resulted in a biased high rate of target points and an incorrect estimate about the regions. Since such regions' photo count is far less than that of the most likely cluster, they have a smaller LLR than most of the true clusters found in the scan statistic process's earlier iterations and could be eliminated by applying a stricter significance level. However, an extremely strict α value increases the risk of excluding a true cluster with a relatively low LLR. The step of removing photo redundancy in Section 3.3.1 focuses only on same-user photos with an identical geolocation but not on same-user photos scattered across a wide area. The phenomenon that a small proportion of Flickr users contributed the majority of geotagged photos is common across the entire study area. This phenomenon does not cause problems in regions with adequate users or regions with few users but with correctly tagged photos. Only those regions with a few people and with wrongly tagged photos tend to be incorrectly estimated because of photo redundancy. The cluster detection itself can hardly handle this bias in VGI data, so a test of whether a detected place extent is supported by enough users is proposed in Section 3.3.3 to help remove false-positive clusters.

Even when using a strict significance level α of 0.001, almost all disjoint extents were still successfully detected. This finding suggests that bursts of photos tagged with a target place name are dominant at the target place. In other words, Flickr users strongly tend to tag a photo with its place name, thus making geotagged photos a good data source for representing places.

Figure 3.7 shows identified significant clusters (blue circles) of disjoint extents compared with reference boundaries. Most detected clusters are centered near the official boundaries' centroids, indicating a good approximation of place locations. Away from the reference boundaries are many outliers possibly caused by users' incorrect knowledge of a place or mistaken tagging behavior. Some of them form clusters at very popular locations distant from target places, such as the one near the Empire State Building in Manhattan and the one at the Golden Gate Bridge in California (see areas D and E in Figure 3.7). The detection process successfully distinguished areas with dense outliers from true clusters of disjoint extents by testing against a random process of the occurrence of photos tagged with a target place name. The successful detection based on this approach may imply a random nature of the outliers located far from target places and scattered over the study area. Successfully removing outlier clusters also proves the importance of considering Flickr photos' uneven spatial distribution.

3.4.2 Results of Outlier Removal in Local Area

Each local study area defined in Section 3.3.3 was separately analyzed using both the RW-KDE outlier removal approach and the modified approach based on the

highest LLR (HLLR). Using the latter, Figure 3.8 shows successful separation of valid points (i.e., the target points estimated not to be outliers) and outliers that could not be identified by the cluster-detection step using the circular scanning window. To quantify performance, we used the traditional measure of accuracy to assess how close the estimated set of valid points (E) is to the real set of photo points within a target place (R). The measure of accuracy is defined as the proportion of correctly estimated valid points ($E \cap R$) to the total number of involved target points ($E \cup R$). As shown in Table 3.2, the HLLR-based approach produced slightly higher accuracies than the original one. Although the accuracy increase is small, the importance of improved inclusion or exclusion of certain points is noticeable in Figure 3.9. The HLLR-based approach correctly kept the target points in unpopular locations (see area A in Figure 3.9), while the original approach identified them as outliers. These points are important for preventing unpopular locations from being excluded from the derived extents. The HLLR-based approach also correctly identified outliers wrongly included in the original approach's result (see area B in Figure 3.9). This identification can help generate more accurate density surfaces in later estimations. The HLLR rule for eliminating outliers is more robust than the rule of keeping at least 95% of the target points. Although the 95% rule is applicable in most cases, sometimes it is not optimal for certain regions such as our two examples.

Table 3.2 Accuracy comparison of outlier removal in local area.

	National Park	Chinatown	Square Park
Original RW-KDE	0.95	0.96	0.83
HLLR-based	0.96	0.96	0.84

3.4.3 Derived Vague Extents

Each disjoint extent was estimated as an independent place to derive a RW-KDE-FDE surface by following the steps of generating the RW-KDE surface: constructing grid-cell locations, calculating point representativeness, incorporating representativeness into the kernel density estimation, and normalizing the density surface to [0, 1]. Figure 3.10 displays each place name's derived surfaces as a combination of individually estimated disjoint vague extents with smoothing bandwidths about 1/2 of the valid points' minimum bounding rectangle's width in each local study area. Other bandwidths of 1/4 and 3/4 were also tested and had no impact on our conclusions.

We estimated the disjoint extents individually because of their potentially different sizes and photo distributions. For example, the National Park's largest disjoint extent in Figure 3.10(a) is over 100 times larger than the smallest one with the same place name. A good smoothing bandwidth for the larger extent could over smooth the smaller one. Also, in the case of National Park (Figure 3.10(a)), Flickr photos are more sparsely distributed in the largest disjoint extent (Death Valley

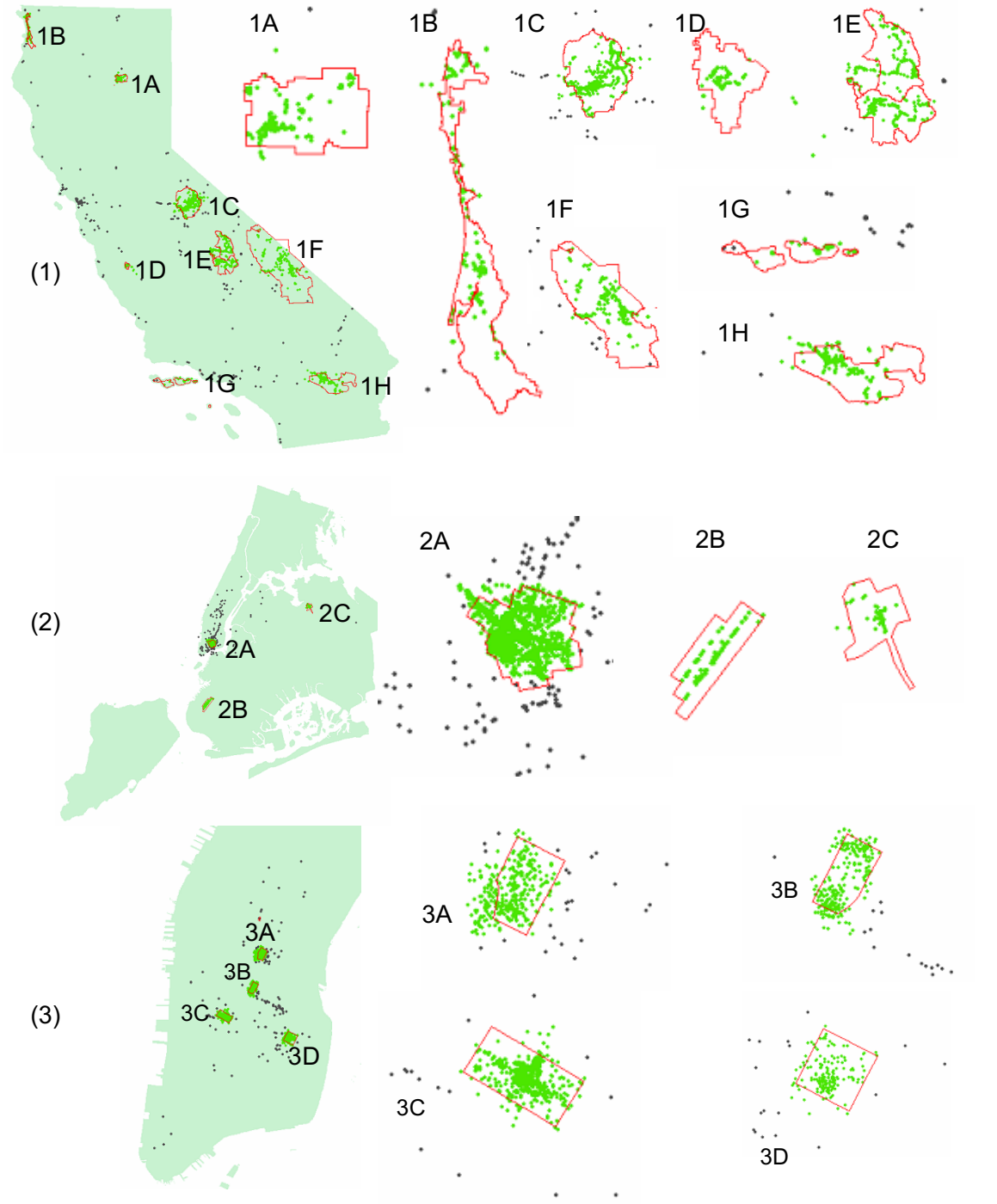


Figure 3.8 Results of outlier removal based on the highest LLR approach in local study areas. Green dots are valid points after outlier removal. Grey dots are outliers to be removed.

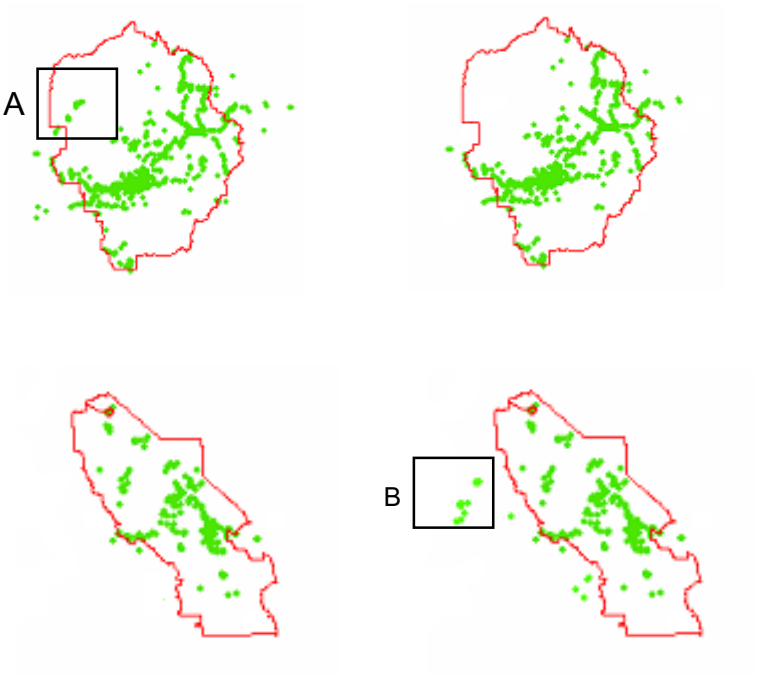
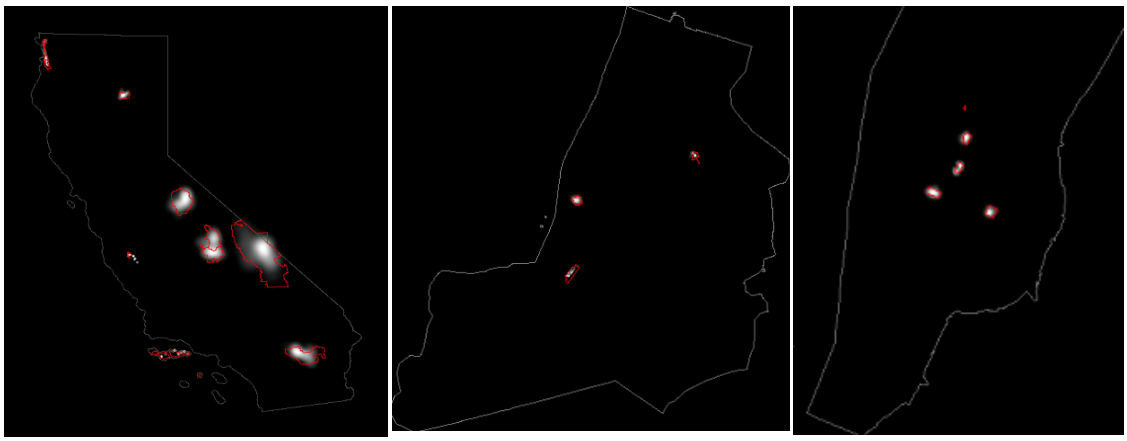


Figure 3.9 Examples of improved outlier removal results. Left: HLLR-based outlier removal. Right: original RW-KDE outlier removal.

National Park) than in the long and narrow extent (Redwood National Park) located in California's northwestern corner. When the grid-cell size for representativeness calculation is chosen, the former extent's suitable size is about four times larger than the latter's appropriate size. Using the former extent's cell size to estimate the latter extent could decrease the representativeness variation in the latter extent and underestimate the photos' representativeness.



(a) (b) (c)

Figure 3.10 RW-KDE-FDE surfaces of (a) National Park, (b) Chinatown, and (c) Square Park.

Traditional KDE and original RW-KDE methods cannot determine disjoint extents or remove outliers for the three selected places. Without dividing the global study area into local study areas, the traditional KDE's and the RW-KDE's parameters must be selected based on the global study area. Figure 3.11 shows an example of the RW-KDE-FDE method's improvement over the traditional KDE and original RW-KDE methods in representing a place's disjoint extents. To make these three methods comparable, we derived their surfaces using the same bandwidth and the same set of valid points that the RW-KDE-FDE method estimated. Figure 3.11(a) indicates the traditional KDE's strong disadvantage in handling location popularity's variation. The extent in Manhattan was estimated to be about 20 times more likely to be Chinatown than the extents located in Brooklyn and Queens. In fact, these three extents are truly Chinatowns and should have similar probability estimates. The original RW-KDE considered photo representativeness and better estimated Brooklyn's Chinatown, but still underestimated the one in Queens. It is interesting to find that although there is a burst of photos tagged as Chinatown at the Queens Chinatown, Flickr users are less likely to tag their photos as Chinatown in this region than in the other two regions. This finding indicates that users' tagging behavior varies across different geographic areas although these areas share the same name. As shown in Figure 3.11(c), RW-KDE-FDE created a much better representation of the three disjoint extents, proving the importance of determining

disjoint extents and estimating each disjoint extent in the local study area.

(a) (b) (c)

Figure 3.11 Density surface comparison of Chinatown: (a) traditional KDE, (b) RW-KDE based on global study area, and (c) RW-KDE-FDE based on local study area.

3.5 Conclusion

Flickr geotagged photos have been regarded as a valuable data source for exploring human perception and knowledge of places. Deriving the spatial extent of places that do not have a formally defined boundary is an important aspect of studying the concept of place. Challenges arise when using geotagged photos to quantify the spatial extent of places, including biased spatial coverage of Flickr photos and data-quality uncertainty. Focusing on these challenges, this paper makes two major contributions.

First, we filled the research gap by providing an effective approach to reducing uneven photo distribution's impact in order to properly derive the extent of a place with disjoint regions. Using the proposed RW-KDE-FDE method, we successfully determined the number and locations of disjoint regions sharing the same place name. Although some regions have a high density of outliers (such as the example in Figure 3.2) they were successfully distinguished from the real clusters representing the disjoint extents. Unlike other cluster-detection approaches based purely on the closeness between observations, the scan-statistic-based method considered the underlying Flickr photo coverage and correctly excluded the outlier clusters from the results. The comparison between the derived RW-KDE-FDE surface and the traditional KDE surface confirms that considering the location's popularity can greatly improve the place extent's estimation based on Flickr data. The comparison to the original RW-KDE method shows the importance and feasibility of addressing a place's disjoint extents.

Second, we explored the scan statistic approach's suitability for Flickr geotagged photos as a type of VGI datasets, which tend to have more complex noises than

actively collected datasets, such as disease records and crime reports. The scan-statistic-based burst-detection approach has been widely used in the latter data type, but rarely applied to VGI datasets. Our results show that by comparing with the simulations of random process the scan statistic approach can reduce the impact of the random errors Flickr users produced. However, this approach has difficulty dealing with the problem caused by a small proportion of Flickr users' uploading most of the Flickr photos. If a frequent user contributes many photos with the wrong information in an area where few other users have uploaded photos, this location's estimate would be biased toward the wrong information. The influence of frequent users' errors can be small in locations where an adequate number of other users have accurately contributed. However, for unpopular locations, more place-related information and criteria other than photo geolocations alone should be considered to improve the approximation of place extents.

The RW-KDE-FDE method produced good surfaces representing the disjoint spatial extents of places, but it has room for improvements. The scan statistic process uses a circular scanning window to search for the maximum likelihood cluster, while the place extent's real shape is irregular. Although the outlier removal process in the local study area found an irregularly shaped cluster, the Delaunay-triangulation-based cluster construction did not consider underlying photo distribution. Thus, the resulting major cluster with maximum LLR may not be the optimal solution. Future work can focus on developing a better scan statistic process, which can find the maximum LLR cluster with an irregular shape.

References

Agnew, J.A., 2011. Space and place. *In*: J.A. Agnew and D.N. Livingstone, eds. *The SAGE handbook of geographical knowledge*. Thousand Oaks, CA: SAGE, 316–330.

Arampatzis, A., et al., 2006. Web-based delineation of imprecise regions. *Computers, Environment and Urban Systems*, 30(4), 436-459.

Casey, E.S., 1997. *The fate of place: a philosophical history*. Berkeley and Los Angeles: University of California Press.

Chen, J. and Shaw, S.L., 2016. Representing the spatial extent of places based on Flickr photos with a representativeness-weighted kernel density estimation. *In*: J.A. Miller, D. O'Sullivan, and N. Wiegand, eds. *Geographic Information Science. Lecture notes in computer science vol. 9927*. Cham: Springer, 130-144.

Cunha, E. and Martins, B., 2014. Using one-class classifiers and multiple kernel learning for defining imprecise geographic regions. *International Journal of Geographical Information Science*, 28(11), 2220-2241.

Dwass, M., 1957. Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics*, 28(1), 181-187.

Dilo, A., De By, R.A., and Stein, A., 2007. A system of types and operators for handling vague spatial objects. *International Journal of Geographical Information Science*, 21(4), 397–426.

Egenhofer, M.J. and Mark, D.M., 1995. Naïve Geography. In: A.U. Frank, W. Kuhn, eds. *Spatial information theory a theoretical basis for GIS. Lecture notes in computer science vol. 988*. Berlin: Springer, 1–15.

Eldershaw, C. and Hegland, M., 1997. Cluster analysis using triangulation. *In*: *Computational techniques and applications, CTAC 97*. Singapore: World Scientific, 201-208.

Ester, M., et al., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *In*: E. Simoudis, J. Han, and U.M. Fayyad, eds. *Proceedings of the second international conference on knowledge discovery and data mining (KDD-96)*. Palo Alto, CA: AAAI Press, 226–231.

Feick, R. and Robertson, C., 2015a. Identifying locally- and globally-distinctive urban place descriptors from heterogeneous user-generated content. In: F. Harvey and Y. Leung, eds. *Advances in spatial data handling and analysis*. Berlin: Springer, 51-63.

Feick, R. and Robertson, C., 2015b. A multi-scale approach to exploring urban places in geotagged photographs. *Computers, Environment and Urban Systems,* 53, 96-109.

Goodchild, M.F., 2011. Formalizing place in geographic information systems. *In*: L.M. Burton, et al., eds. *Communities, neighborhoods, and health*. New York: Springer, 21-33.

Goodchild, M.F., 2015. Space, place and health. *Annals of GIS*, 21(2), 97-100.

Goodchild, M.F., Aubrecht, C., and Bhaduri, B., 2016. New questions and a changing focus in advanced VGI research. *Transactions in GIS*, 1-2.

Goodchild, M.F. and Li, L., 2012. Assuring the quality of volunteered geographic information. *Spatial Statistics*, 1, 110-120.

Grothe, C. and Schaab, J., 2009. Automated footprint generation from geotags with kernel density estimation and support vector machines. *Spatial Cognition & Computation*, 9(3), 195-211.

Guo, Q., Liu, Y., and Wieczorek, J., 2008. Georeferencing locality descriptions and computing associated uncertainty using a probabilistic approach. *International Journal of Geographical Information Science*, 22(10), 1067-1090.

Hollenstein, L. and Purves, R., 2010. Exploring place through user-generated content: using Flickr tags to describe city cores. *Journal of Spatial Information Science*, 1, 21-48.

Hu, Y., et al., 2015. Extracting and understanding urban areas of interest using geotagged photos. *Computers, Environment and Urban Systems*, 54, 240-254.

Jones, C.B., et al., 2008. Modelling vague places with knowledge from the web. *International Journal of Geographical Information Science*, 22(10), 1045–1065.

Kulldorff, M.A., 1997. Spatial scan statistic. *Communications in Statistics: Theory and Methods*, 26(6), 1481-1496.

Kulldorff, M., et al., 2006. An elliptic spatial scan statistic. *Statistics in Medicine*, 25(22), 3929–3943.

Kulldorff, M., 2015. *SaTScan™ user guide for version 9.4*. Available from: http://www.satscan.org/cgi-bin/satscan/register.pl/SaTScan_Users_Guide.pdf?todo=process_userguide_download. [Accessed October 1, 2016].

Kulldorff, M. and Information Management Services Inc., 2015. *SaTScan™ v9.4: Softeware for the spatial and space-time scan statistics*. http://www.satscan.org/. [Accessed May 1, 2016].

Li, L. and Goodchild, M.F., 2012. Constructing places from spatial footprints. *In*: *Proceedings of the 1st ACM SIGSPATIAL international workshop on crowdsourced and volunteered geographic information*. New York: ACM, 15-21.

Lüscher, P. and Weibel, R., 2013. Exploiting empirical knowledge for automatic delineation of city centres from large-scale topographic databases. *Computers, Environment and Urban Systems*, 37, 18-34.

Martins, B., 2011. Delimiting imprecise regions with georeferenced photos and land coverage data. *In*: K. Tanaka, P. Fröhlich, and K. Kim, eds. *Web and wireless geographical information systems. Lecture notes in computer science vol. 6574*. Berlin: Springer, 219-229.

Mackaness, W.A. and Chaudhry, O., 2013. Assessing the veracity of methods for extracting place semantics from Flickr tags. *Transactions in GIS*, 17(4), 544-562.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. *In*: L.M. Le Cam and J. Neyman, eds. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Volume 1: statistics*. Berkeley, CA: University of California Press, 281–297.

McKenzie, G., et al., 2015. How where is when? On the regional variability and resolution of geosocial temporal signatures for points of interest. *Computers, Environment and Urban Systems*, 54, 336-346.

Montello, D.R., et al., 2003. Where's Downtown?: Behavioral methods for determining referents of vague spatial queries. *Spatial Cognition & Computation*, 3(2-3), 185–204.

Montello, D.R., Friedman, A., and Phillips, D.W., 2014. Vague cognitive regions in geography and geographic information science. *International Journal of Geographical Information Science*, 28(9), 1802-1820.

O'Hare, N. and Murdock, V., 2013. Modeling locations with social media. *Information Retrieval*, 16(1), 30-62.

Openshaw, S., et al., 1987. A mark I geographical analysis machine for the automated analysis of point data sets. *International Journal of Geographical Information Science*, 1(4), 335–358.

Purves, R.S. and Derungs, C., 2015. From space to place: place-based explorations of text. *International Journal of Humanities and Arts Computing*, 9(1), 74-94.

Rattenbury, T. and Naaman, M., 2009. Methods for extracting place semantics from Flickr tags. *ACM Transactions on the Web (TWEB)*, 3(1), 1-30.

Schockaert, S., De Cock, M., and Kerre, E.E., 2005. Automatic acquisition of fuzzy footprints. In: R. Meersman, Z. Tari, and P. Herrero, eds. *On the move to meaningful internet systems 2005: OTM 2005 workshops. Lecture notes in computer science vol. 3762*. Berlin: Springer, 1077-1086.

Sui, D. and Goodchild, M., 2011. The convergence of GIS and social media: challenges for GIScience. *International Journal of Geographical Information Science*, 25(11), 1737-1748.

Tuan, Y.-F., 1977. *Space and place: the perspective of experience*. Minneapolis: University of Minnesota Press.

Chapter 4

Where and When Taxi Drivers Deviate from the Shortest-Distance Routes in A City

A version of this chapter's Subsection 4.4.1 was presented at GIScience 2012 by Jiaoli Chen, Shih-Lung Shaw, Yang Xu, Qingquan Li, Zhixiang Fang and Yuguang Li:

Chen, J., et al., 2012. Where and when taxi drivers deviate from the shortest path in their route choices: A case study of Wuhan, China. *In*: N. Xiao, M.-P. Kwan, and H. Lin, eds. *GIScience 2012 extended abstracts: proceedings*. http:// http://www.giscience.org/past/2012/proceedings/abstracts/giscience2012_paper_90.pdf.

This chapter substantially extended and revised the original work. The datasets of taxi GPS trajectories and road network were provided by the last three coauthors. The research analyses were made by Jiaoli Chen advised by Dr. Shih-Lung Shaw. Yang Xu provided suggestions during this research.

Abstract

Understanding how drivers choose routes can not only benefit transportation policy making, but also enhance navigation systems based on drivers' experience. This paper uses a large volume of taxi GPS (Global Positioning System) tracking data collected in Wuhan, China to uncover taxi drivers' route choice patterns in different situations by comparing actual routes to the shortest-distance routes. Two indices are proposed to measure taxi drivers' preference for and avoidance of each road segment. The indices are used to uncover the spatiotemporal patterns of taxi drivers' deviations from the shortest-distance routes. By analyzing the relationships between the patterns and the road functional class, travel distance, and urban rush hours, this paper finds that deviation from the shortest-distance route is influenced more by road functional class and travel distance than by urban rush hours. Taxi drivers tend to use the shortest-distance routes on high-hierarchy roads and short trips. In contrast, they tend to frequently deviate from the shortest-distance routes in areas with a high density of local streets. Furthermore, they are not likely to choose a primary road as an alternative if it is not on the shortest-distance route. However, once a primary road is on the shortest-distance route, drivers tend to stay on it regardless of time-varying traffic conditions. Interestingly, on most of the urban roads in Wuhan, China, the difference in deviation rate between rush hours and off-rush hours is small. This finding may suggest that taxi drivers' route-choice behavior could be more consistent than what is usually expected when the most urban roads' traffic conditions get worse during rush hours.

4.1 Introduction

Route choice decisions are frequently made in daily life and greatly impact travel efficiency. Many route choice models and navigation systems use the shortest routes as the default assumption when describing route-choice behavior and providing directions because such routes are "simple, intuitive, and easy to

implement" (Levinson and Zhu 2013, p. 232). However, real-world route decisions are complex and involve many factors such as route attributes, traveler's sociodemographic characteristics (see Jan et al. 2000), and "social and normative contexts" (Spissu et al. 2011, p. 96). The shortest route may not reflect the actual route that the driver regards as the best option. Understanding how drivers' actual routes are different from the shortest routes can not only inform transportation modeling and policy making, but also enhance navigation systems based on drivers' experience.

Great efforts have been made to improve route choice models (e.g., Cascetta et al. 1996, Ben-Akiva and Bierlaire 1999) for better representation of route decisions. However, the models are often built on very limited information about the actual routes (Patro 2009). With GPS' increasing popularity, collecting data of actual routes becomes relatively easier, thus promoting many empirical studies. Because of the constraint created by the tedious data-collection process, much work examining the difference between actual routes and the shortest routes (e.g., Papinski and Scott 2013) is based on small samples collected from only a few drivers. From such data, it is hard to acquire information about what route decisions most drivers would make at certain locations and times and in certain situations. In recent years, the availability of taxi GPS tracking data has made it possible to obtain large volumes of taxi drivers' actual routes, which could cover almost an entire city and long time periods. Through frequent routing practices during work, taxi drivers tend to accumulate experience and become familiar with the urban road network's dynamic traffic conditions (Li et al. 2011). Thus, it becomes possible to reveal the route-choice patterns of taxi drivers in different situations and to obtain their experience and knowledge.

Drivers' route choice decisions can vary across place and time. For example, in the real world, one would follow the shortest route on certain roads but might deviate from it at another place in the trip. Also, some roads are preferred by many drivers at certain times of the day but are avoided by most drivers at other times because of heavy congestion or other reasons. This paper aims to explore the spatiotemporal patterns of taxi drivers' route-choice behavior by examining how likely those drivers would be to prefer or avoid a road segment in different situations (e.g., travel distance, time window and road functional class) compared with the shortest-distance route. For each road segment, such information can be stored in a geographic information system (GIS) as a reference for future decision-making processes. By examining different factors' roles in the deviation patterns, we also aim to determine what unknown patterns can be found from taxi tracking data and whether the conclusions drawn from such big data are consistent with the conclusions drawn from traditional survey and sample data.

The remainder of this article is organized as follows. Section 4.2 reviews work related to route choice behavior. Section 4.3 describes the datasets. Section 4.4

includes methods, results and discussions. Subsection 4.4.1 proposes a road-segment-based approach to measure how likely taxi drivers are to prefer or avoid a road segment and the temporal variation. City roads are categorized into two sets of four groups with each representing the locations having similar deviation patterns. How road functional class is related to the patterns is also examined. Subsection 4.4.2 uses a trip-based approach to answer how travel distance influences drivers' road class preference in actual routes compared to the shortest-distance routes. Subsection 4.4.3 summarizes taxi drivers' route-choice behavior under different situations by combining long/short travel distances, peak/off-peak hours, and road classes. Lastly, Section 4.5 discusses implications, limitations, and future research.

4.2 Related Work

There has been a long history of examining and simulating human routing behavior in many research fields such as transportation engineering, GIS and computer science. Most methods make assumptions based on little information about the actual routes and have static views when representing human's complex route-choice behavior. The validity of the assumptions and the representation performance are questionable. For example, the commonly-used deterministic user equilibrium model in transportation engineering assumes that drivers are knowledgeable about the optimal route and always take the minimum-cost route according to Wardrop's (1952) first principle (Prato 2009). However, an optimal route in distance, travel time or expense is not necessarily optimal for people who might have different feelings, preferences and knowledge. Although the framework of discrete choice model (e.g., Daganzo and Sheffi 1977, Frejinger and Bierlaire 2007) better reflects drivers' "nonoptimal behavior" (Dial 1971, p. 83), it usually assumes that route choice does not change over time. Some agent-based models assume that drivers can dynamically react to the changes in traffic conditions and can modify decisions en route (e.g., Rossetti et al. 2000, Dia 2002). However, these models are usually based on theoretical calculations and do not incorporate detailed information about actual route decisions. Understanding how actual routes are different from theoretical assumptions and harvesting information about route choice behavior can help improve model assumptions and simulation methods.

Several studies have applied actual route data (e.g., stated and revealed preference surveys and GPS observations) to empirically estimate existing models, determine factors influencing route choice decisions, and explore route choice patterns. For example, some studies focused on estimating existing route-choice models' goodness-of-fit with the actual route data (e.g., Bekhor et al. 2006, Frejinger and Bierlaire 2007, Prato et al. 2012). However, they did not focus on how and why the actual routes were different from those estimated from the models. Some studies aimed to determine route choice factors, such as the trips'

and the routes' attributes (Dalton 2003) as well as the driver's sociodemographic characteristics, perception and cognition of route (Parkany et al. 2006, Tawfik et al. 2010), and use of traffic information (Abdel-Aty et al. 1997, Chen et al. 2001, Zhang and Levinson 2008). For example, Li et al. (2005) used 182 drivers' morning commute routes collected by GPS devices to build a binary logit model, which could predict drivers' choices between single versus multiple commute routes. The authors identified the most influential factors to be driver age, income, and flexible work schedule. Another set of studies focused on exploring actual routes' patterns by examining the diversions from the theoretical assumptions (e.g., Papinski and Scott 2011), route choice variations among individuals (e.g., Spissu et al. 2011), differences between observed routes and planned routes (e.g., Papinski et al. 2009), to name a few. Pattern exploration is a good way to discover unknown features, which could inform future research exploring better route choice models. This study follows this direction to explore taxi drivers' deviation patterns from the shortest-distance routes. All the above-mentioned empirical studies are based on surveyed and sampled data that cover a small study area and short time span and that represent only a few drivers. Thus, it is not easy to identify unknown patterns from such data pertaining to the spatial and temporal variations of route choice behavior.

Some studies examined how actual routes are different from the shortest routes, but did not address spatiotemporal variations. For example, Jan et al. (2000) compared GPS routes to their counterpart shortest-time routes and observed important differences in travel time and roads taken. Zhu and Levinson (2010) followed up to address the extent to which the actual routes are different from the shortest-time routes among commute trips and non-commute trips. Papinski and Scott (2011) developed a GIS-based toolkit to facilitate the comparison between a GPS route and its corresponding shortest-time and shortest-distance routes in terms of route attributes (e.g., travel time, number of intersections and route directness). The results imply that real-world route choices cannot be reflected by the shortest-distance route assumption. They did not provide further information about where and when the deviations happened. Only a few studies have included spatial or temporal considerations. For example, Jan et al. (2000) noted that the paths chosen by the same driver for the same trip at different times were consistent. Ramaekers et al. (2013) found that trips in different geographic regions vary greatly in the degree of deviation from the shortest route. Thus, more efforts are needed to extend these studies to address the spatiotemporal variation on a larger scale and with higher spatial and temporal resolutions. New methods that can better facilitate spatiotemporal analysis of route choice behavior are needed.

In this era of big data, large volumes of GPS traces are broadly and frequently applied in various research topics including human mobility patterns (e.g., Giannotti et al. 2011), dynamic traffic condition (e.g., Ehmke et al. 2012), and land use extraction (e.g., Liu et al. 2012). However, the number of studies that take

advantage of big trajectory data to analyze route choice behavior is limited. For example, Liu et al. (2010) used a large number of GPS traces to analyze the difference in route choice patterns between two taxi driver groups of different income levels. However, they did not focus on acquiring information about how taxi drivers choose their routes under certain situation and at different locations and times. Filling this research gap not only helps gain new insights into route choice behavior but also informs future navigation applications.

4.3 Data

For our study, a GPS trajectory dataset of more than 11,000 taxis was continuously collected from Wuhan, China, during one week in March 2009. This dataset was preprocessed to eliminate any GPS data irregularities and to match the trajectories to approximately 45,000 road segments taking travel direction into consideration. Individual passenger trips and their durations were then identified from these trajectories based on changes in passenger status. A *passenger trip* is defined as a trip from a passenger's origin to his/her destination. The GPS trajectory of each passenger trip is associated with a sequence of road segments, representing the actual route a taxi driver takes.

Only passenger trips were analyzed because taxi drivers may practice different route choice behaviors when they are heading to a specific destination versus when they are searching for customers. Finally, more than 440,000 passenger trips per day during the study week were analyzed. Note that the term *trip* to refer to *passenger trip* is used in the remainder of this paper. We performed the same analysis for each day of the study week. Wuhan's street network is an important reference for interpreting the results in the following sections. It includes six functional classes: freeway, state road, city expressway, urban major arterial, urban minor arterial, and local street (see Table 4.1 and Figure 4.1). The first four functional classes are often considered to be at a higher hierarchical level of roads than the last two.

4.4 Methods and Results

4.4.1 Where and When Taxi Drivers Do Not Choose the Shortest-Distance Routes

4.4.1.1 Measures of Deviations from the Shortest-Distance Routes on a Road Segment

For each road segment, we developed two indices (from two complementary perspectives), which quantify the degree of taxi drivers' deviation from the shortest-distance routes. Because of various factors (e.g., traffic conditions and driver knowledge), some road segments at certain times of day might frequently be

Table 4.1 Functional classification of Wuhan road network.

Functional Class	Description
Freeway	Includes inter-province and inter-city highways surrounding the urban area and an airport highway to the northwest of Wuhan
State road	Includes both a major beltway passing through the suburban area and several major roads connecting urban and suburban areas
City expressway	Includes the most important urban arterials and the Wuhan Yangtze River Bridge, connecting Wuhan's three towns (Hankou, Hanyang, and Wuchang) separated by two rivers
Urban major arterial	Includes the second-most important roads and the Second Wuhan Yangtze River Bridge, connecting city expressways with urban minor arterials
Urban minor arterial	Contains the most significant share of road segments in the city, connecting major roads with local streets
Local street	Consists of streets providing access to properties

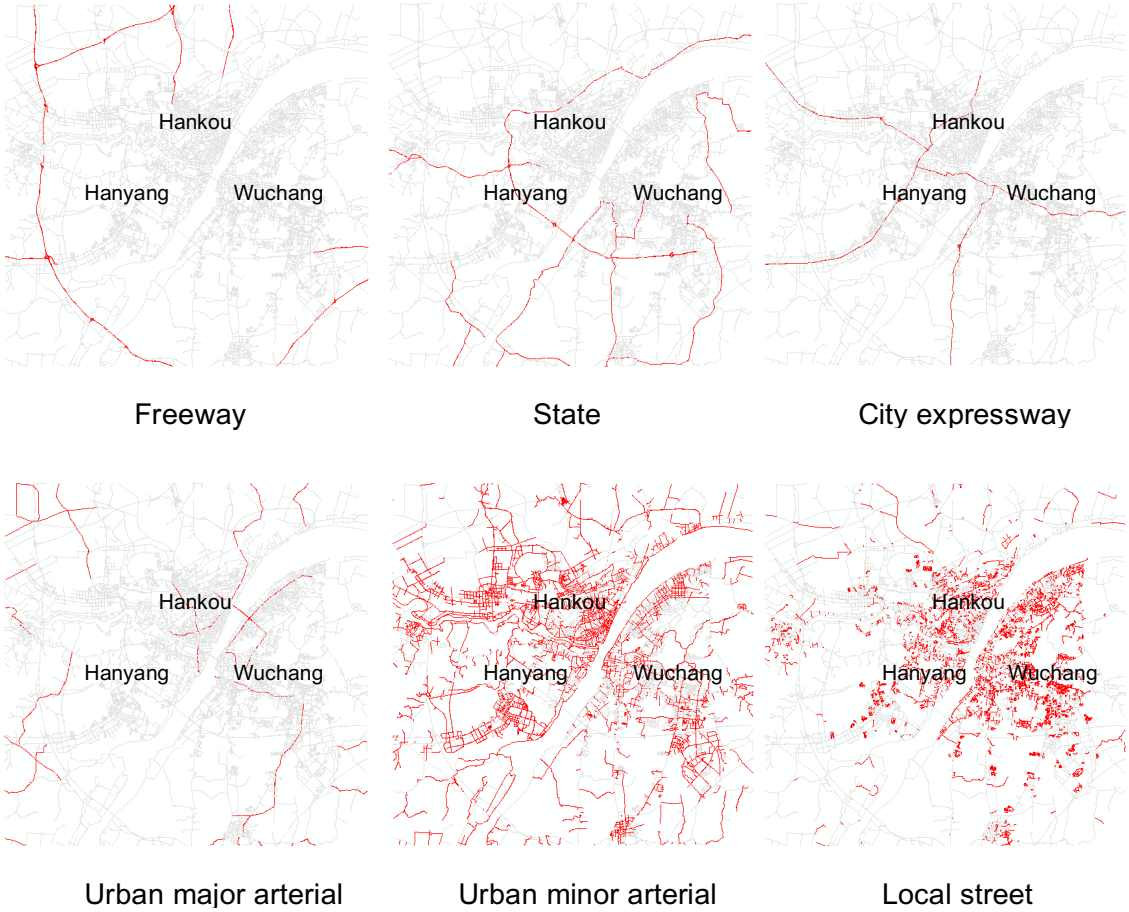


Figure 4.1 Spatial distributions of six functional classes of Wuhan road network.

chosen by taxi drivers even though those segments are not on the shortest-distance routes. Conversely, some road segments at certain times of day might frequently be avoided by taxi drivers even though taking those segments could shorten routes. To examine where and when these cases occur, we first calculated the shortest-distance route for each of the approximately 440,000 trips per day during the study week and associated that route with the corresponding road segments. A road segment and a trip are mutually related only when the road segment is on either or both of the GPS-Tracking Route (GTR) (i.e., the actual route taken) and the Shortest-Distance Route (SDR) of the trip. If related to one of the trips, a road segment must be in one of three scenarios:

(a) *Matched road segment*: a road segment that is chosen by a taxi driver *and* is on the trip's shortest-distance route (Figure 4.2 (a));

(b) *NonSDR road segment*: a road segment that is chosen by a taxi driver but is *not* on the trip's shortest-distance route (Figure 4.2 (b));

(c) *NonGTR road segment*: a road segment that is not chosen by the taxi driver but is on the trip's shortest-distance route (Figure 4.2 (c)).

Each day's trips were further divided into 48 half-hour windows based on start and end times. The half-hour window was chosen to make sure both urban and suburban areas' road segments have sufficient trips in each time window for calculating indices. If a trip's duration crosses the time-window boundary, the trip was assigned to the time window that has a larger overlap with the trip duration.

We then counted the frequency with which each road segment is classified as a matched road segment ($N^i_{M(t)}$), a NonSDR road segment ($N^i_{NS(t)}$), or a NonGTR road segment ($N^i_{NG(t)}$) for each time window in a day. This count was based on trips falling into the corresponding time window, where *t* represents a time window and *i* is the road segment ID. Finally, the following indices were calculated to measure taxi drivers' deviation from the shortest-distance routes for road segment *i* within time window *t* on a given day:

(1) *NonSDR Ratio* ($NSR^i_{(t)}$): the ratio of the frequency that the road segment is a NonSDR road segment ($N^i_{NS(t)}$) to the total number of actual routes passing through this road segment (i.e., $N^i_{NS(t)} + N^i_{M(t)}$). For example, a NonSDR ratio of 0.8 states that 80% of taxi trips using this road segment are not on the shortest-distance route.

(2) *NonGTR Ratio* ($NGR^i_{(t)}$): the ratio of the frequency that the road segment is a NonGTR road segment ($N^i_{NG(t)}$) to the total number of the shortest-distance

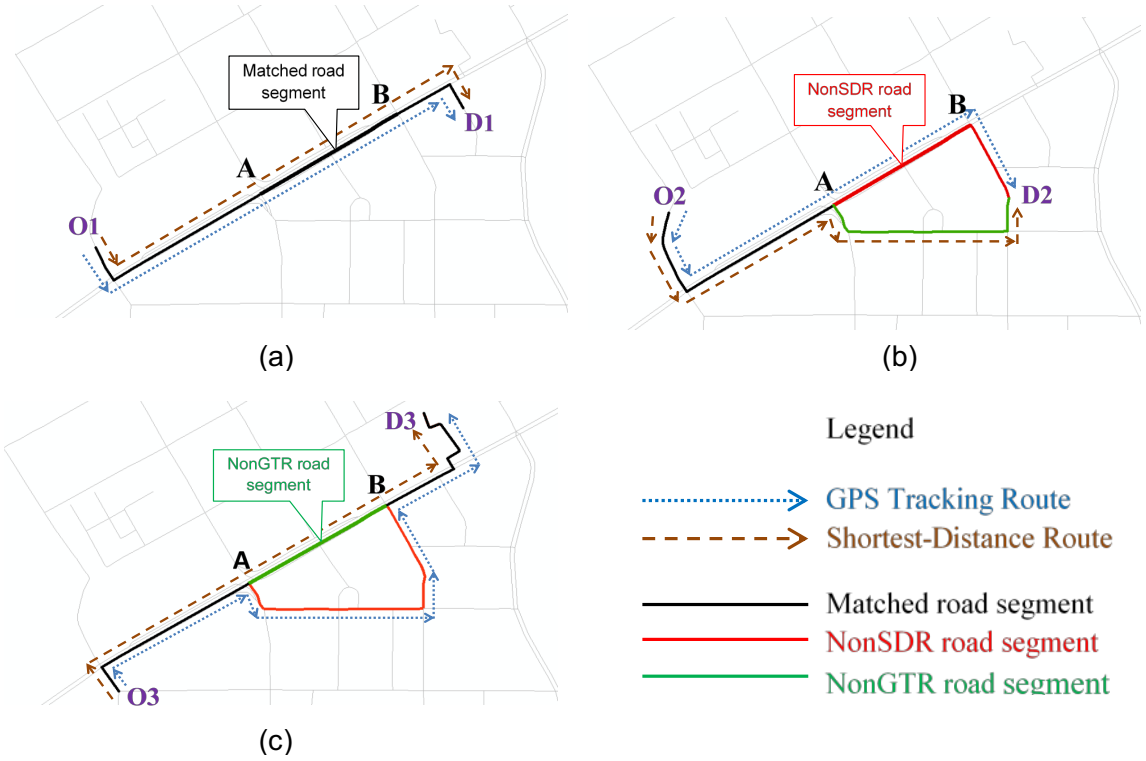


Figure 4.2 Three scenarios for one road segment (from junction A to junction B) related to three different taxi trips (O1 to D1, O2 to D2 and O3 to D3): (a) Matched road segment, (b) NonSDR road segment, and (c) NonGTR road segment.

routes passing through this road segment (i.e., $N^i_{NG(t)} + N^i_{M(t)}$). For example, a NonGTR ratio of 0.2 indicates that 20% of taxi trips do not choose this road even though it is on their shortest routes.

These two indices of a road segment indicate how likely taxi drivers would be to prefer or avoid the segment compared to the shortest-distance route. Thus, this information can be stored in a GIS to be retrieved by future applications. A road segment with a higher NonSDR ratio indicates that drivers are more likely to choose it even though it can increase their travel distances, while a road segment with a higher NonGTR Ratio indicates that drivers tend to avoid it even if route distances would increase.

4.4.1.2 Spatiotemporal Patterns of Deviations from the Shortest-Distance Routes

Taxi drivers could adjust their route decisions according to time-dependent factors; thus, a road segment would have very different values of NonSDR ratios and NonGTR ratios throughout a day. To examine the temporal pattern of drivers' route decisions on each road segment, we first calculated the mean and variance of NonSDR ratios of all 48 time windows and then computed the *variance-to-mean ratio* ($VMR = variance/mean$) for the entire day. The *VMR* is an efficient and effective way of indicating whether a road segment's NonSDR ratio is persistent over time. The *mean* value indicates the NonSDR ratio's average magnitude for a road segment in a day.

Some road segments have a higher mean of NonSDR ratios in a day and are more persistent over the 48 time windows than other road segments. To find which road segments share a similar pattern, we grouped the road segments into different categories. According to all road segments' mean values in a given day, the median (M_m) bisects the means into two levels: high (≥ M_m) and low (< M_m). Also, based on all road segments' VMRs on the same day, the median value (M_{vmr}) separates the road segments with persistent (< M_{vmr}) NonSDR ratios from those with variable (≥ M_{vmr}) NonSDR ratios. Thus, a road segment can be classified into one of the four categories based on its NonSDR ratios's mean and VMR on that particular day:

(1) *Persistently high NonSDR ratios*: *mean* ≥ M_m and *VMR* < M_{vmr};
(2) *Variably high NonSDR ratios*: *mean* ≥ M_m and *VMR* ≥ M_{vmr};
(3) *Persistently low NonSDR ratios*: *mean* < M_m and *VMR* < M_{vmr}; and
(4) *Variably low NonSDR ratios*: *mean* < M_m and *VMR* ≥ M_{vmr}.

We also performed a similar calculation and classification for the NonGTR ratios. A road segment can also be classified into one of the following four categories based its NonGTR ratios' mean and VMR compared to the median (M'_m) of all road segments' means and the median (M'_{vmr}) of all road segments' VMRs:

(1) *Persistently high NonGTR ratios*: *mean* ≥ M'_m and *VMR* < M'_{vmr};
(2) *Variably high NonGTR ratios*: *mean* ≥ M'_m and *VMR* ≥ M'_{vmr};
(3) *Persistently low NonGTR ratios*: *mean* < M'_m and *VMR* < M'_{vmr}; and
(4) *Variably low NonGTR ratios*: *mean* < M'_m and *VMR* ≥ M'_{vmr}.

Table 4.2 elaborates on each category's meaning. Figure 4.3 shows the temporal distributions of the road segments' NonSDR ratios and NonGTR ratios selected as examples from each category. Some time windows do not have observations and are displayed as broken lines. They were also excluded from the calculation of mean and VMR values. We mapped the spatial distribution of each category of NonSDR ratios and NonGTR ratios in Figure 4.4 and Figure 4.5, respectively. The road segments that do not have enough related trips for calculating indices are not classified into any of these categories and displayed in gray. The ratios show two sets of four distinct spatial patterns among the different categories. When the process was repeated for each day of the study week (including week days and the weekend), these patterns are similar. This finding may suggest that the aggregated patterns of taxi drivers' deviation from the shortest-distance routes are relatively stable in the urban road network, although a driver's route choice could be influenced by the dynamic traffic condition at the individual level.

The NonSDR ratios' four categories (Figure 4.4) show similar spatial patterns with the NonGTR ratios' four categories (Figure 4.5). In each set of four categories, the roads sharing similar spatiotemporal patterns of route-choice behavior appear to be related to certain road functional classes. From the perspective of persistently high NonSDR ratios (Figure 4.4(a)), the road segments are mainly short local streets with a high network density. A similar pattern is also found in the category of persistently high NonGTR ratios (Figure 4.5(a)), Regarding the persistently low NonSDR ratios (Fig.4.4((c)), the road segments appear to be primary roads in urban areas, also similar to the persistently low NonGTR ratios (Figure 4.5(c)).

4.4.1.3 Functional Classes and Discovered Patterns

The spatial distributions of the categories in Figure 4.4 and 4.5 appear to have different patterns in functional class hierarchy as described above. To quantitatively examine if functional class is related to the classification of the four categories of NonSDR ratios, we first calculated the functional class composition for each of the four categories: the percentages of road segments belonging to each functional class among road segments within a category. We also calculated the general functional class composition across the four categories: the percentages of road segments belonging to each functional class among all road segments in the four categories. To know if a category has a much higher percentage in certain functional classes, we computed the relative change of each

Table 4.2 Categories of NonSDR ratios and Categories of NonGTR ratios.

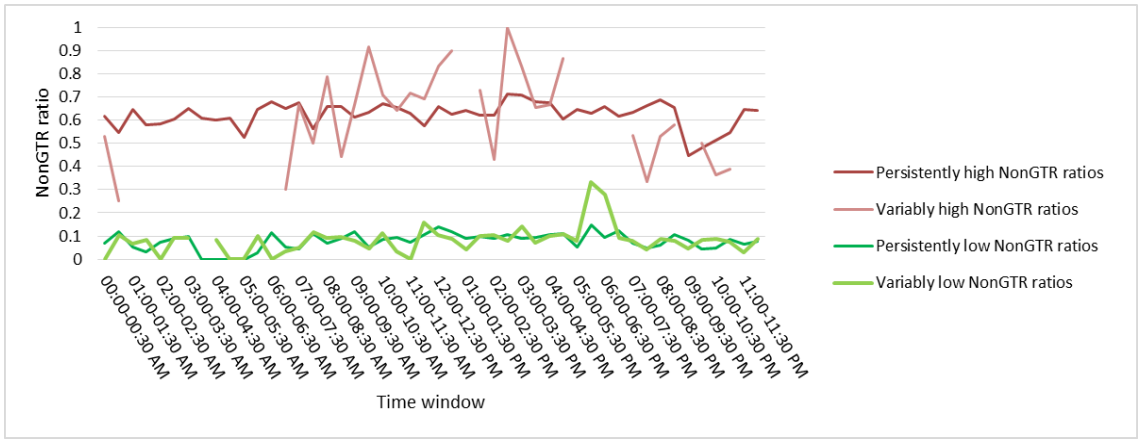
(a) Categories of NonSDR ratios.

	High	Low
Persistently	At any time in a day, a road segment in this category is chosen more by taxi drivers as an alternative road to the shortest-distance route.	At any time in a day, a road segment in this category is chosen less by taxi drivers as an alternative road to the shortest-distance route.
Variably	A road segment in this category could be chosen either more or less by taxi drivers as an alternative road to the shortest-distance route at some time during a day. However, on average it is chosen more by taxi drivers as an alternative road to the shortest-distance route.	A road segment in this category could be chosen either more or less by taxi drivers as an alternative road to the shortest-distance route at some time during a day. However, on average it is chosen less by taxi drivers as an alternative road to the shortest-distance route.

(b) Categories of NonGTR ratios.

	High	Low
Persistently	At any time during a day, a road segment in this category is avoided more by taxi drivers even though it is on their shortest-distance routes.	Whenever a road segment in this category is part of the shortest-distance route, it is chosen more by taxi drivers in their actual routes.
Variably	A road segment in this category could be either avoided more or chosen more by taxi drivers when it is on their shortest-distance routes at some time during a day. However, on average it is avoided more by taxi drivers even though it is on the shortest-distance routes.	A road segment in this category could be either avoided more or chosen more by taxi drivers when it is on their shortest-distance routes at some time during a day. However, on average it is chosen more by taxi drivers when it is part of the shortest-distance routes.

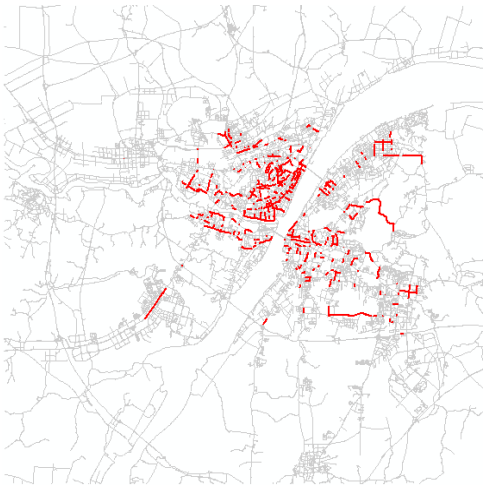
(a)



(b)

Figure 4.3 Temporal distributions of (a) NonSDR ratios and (b) NonGTR ratios of the road segments selected as examples from each category.

(a) Persistently high NonSDR ratios

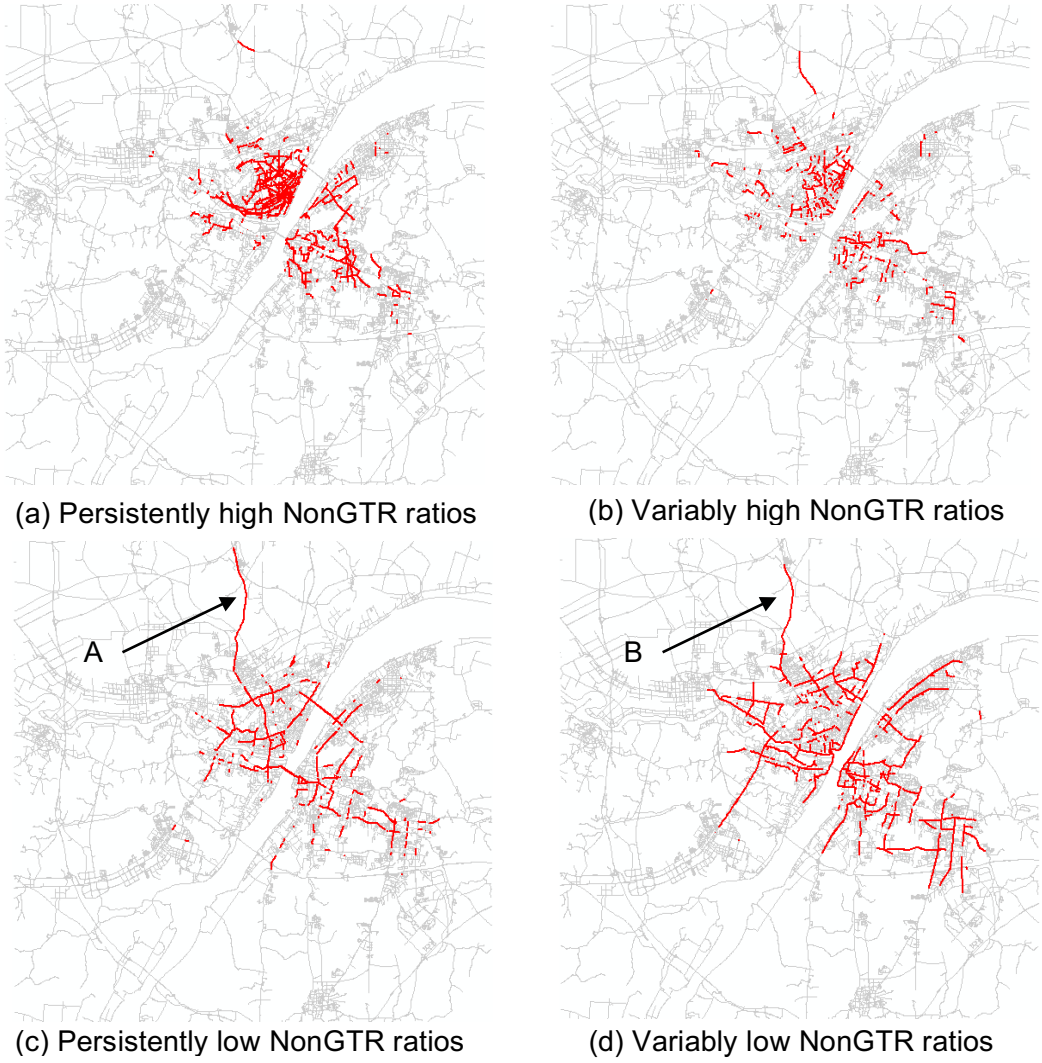


(b) Variably high NonSDR ratios

(c) Persistently low NonSDR ratios

(d) Variably low NonSDR ratios

Figure 4.4 Spatial distributions of the four categories of NonSDR ratios.



(a) Persistently high NonGTR ratios

(b) Variably high NonGTR ratios

(c) Persistently low NonGTR ratios

(d) Variably low NonGTR ratios

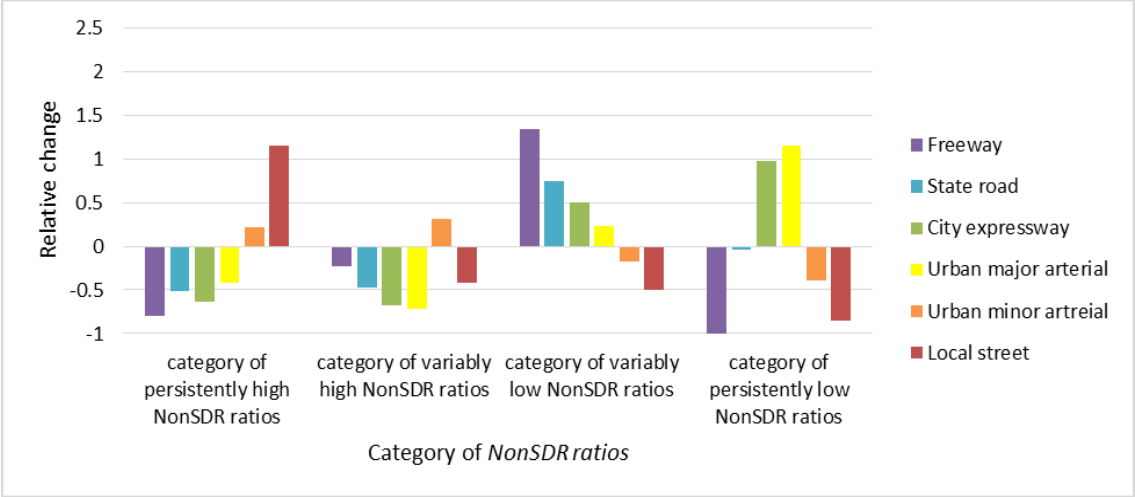
Figure 4.5 Spatial distributions of the four categories of NonGTR ratios.

functional class's percentage (y axis in Figure 4.6) between each category's functional class composition and the general functional class composition (Figure 4.6(a)). If the four categories are not different in terms of functional class, the relative changes would approach a value of 0. The same process was repeated for the NonGTR ratios, and the results are displayed in Figure 4.6(b).

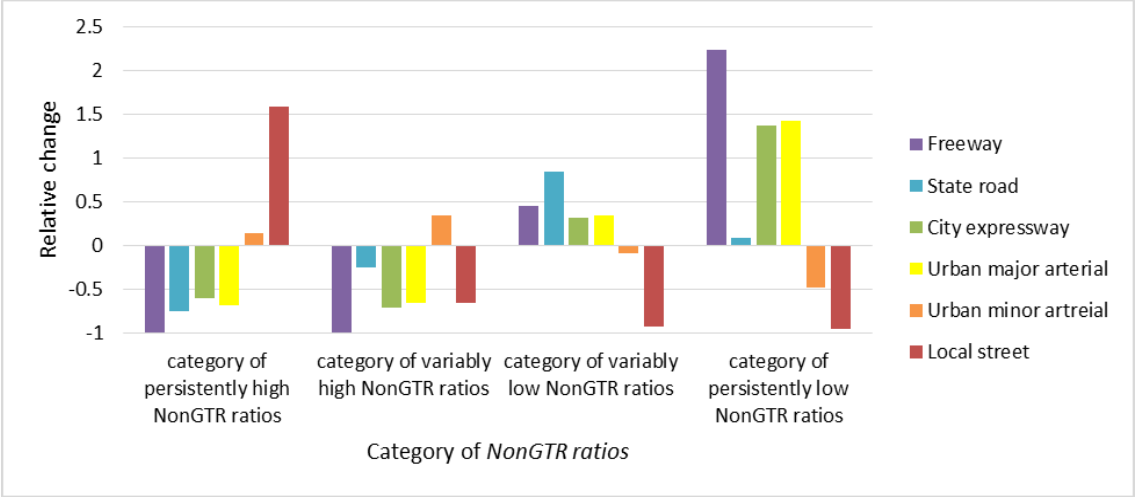
Figure 4.6 (a) and (b) show similar functional class patterns for the four categories between the NonSDR ratios and the NonGTR ratios, except for the freeways in the categories of persistently low NonSDR ratios and persistently low NonGTR ratios. The category of persistently high NonSDR ratios includes roads consistently preferred, while the category of persistently high NonGTR ratios includes roads consistently avoided. Both categories, which reflect opposite route choice behaviors, have a large proportion of low-hierarchy roads including local streets and minor arterials. This pattern indicates that in areas with a high density of local streets and minor roads, taxi drivers may tend to frequently deviate from the shortest-distance route to another nearby local street that probably has better traffic conditions.

The persistently low NonSDR ratios include roads that are not likely chosen if not on the shortest-distance route, while the category of persistently low NonGTR ratios includes roads that are frequently chosen once they are on the shortest-distance route. Both categories, which also reflect opposite route choice behaviors, have a large proportion of city expressways and urban major arterials. In contrast to the above situation in local street areas, taxi drivers are more likely to follow the shortest-distance routes on urban primary roads (i.e., referring urban major arterials and city expressways). They tend not to sacrifice route distance for a primary road that is not on their shortest routes throughout the day, even though traffic on primary roads usually moves at a higher speed if no congestion occurs. Taxi drivers also tend not to give up a primary road on the shortest-distance route at any time of day, even if congestion occurs during rush hours.

Regarding the difference in freeway between persistently low NonSDR ratios and persistently low NonGTR ratios shown in Figure 4.6, the freeways causing the difference are mainly airport highways going toward downtown (see A in Figure 4.5). Taxi drivers tend to consistently use these road segments if they are on the shortest-distance routes. When these roads are not on the shortest-distance routes, whether taxi drivers tend to choose them as alternatives is influenced by time of day. Interestingly, the airport highways in the opposite direction (i.e., going toward the airport) show a different deviation pattern (see A and B in Figure 4.5). The different patterns suggest that when airport highways are on the shortest-distance routes, taxi drivers are more likely to deviate from them at some time during a day in the direction of going toward airport (B) than in the direction of going toward downtown (A). The reason might be that when driving passengers to catch flights, taxi drivers may have concerned more about time cost and tend to



(a)



(b)

Figure 4.6 Comparison of functional class composition among different categories.

detour to a faster route from the shortest-distance route at some time during a day. However, when driving toward downtown, they may have concerned less about time cost and tend to stay on the shortest-distance routes regardless of time of day.

The different route choice behavior on low-hierarchy roads vs. high-hierarchy roads is rarely reported in the literature. Intuitively, we thought drivers would tend to detour to high-hierarchy roads, a conclusion which is also implied in some research (e.g., Li 2004, Ramming 2002, Zhu 2010). However, our finding shows that taxi drivers do not tend to choose urban primary roads as alternatives to the shortest-distance routes. Only when those roads are part of the shortest-distance routes do taxi drivers tend to stay on them. Such behavior is consistent over 48 time windows. This finding possibly indicate that good options on some low-hierarchy roads are available as alternatives and that drivers do not necessarily need to detour to an urban primary road.

The categories with variable ratios show transition in functional class composition between persistently high and persistently low categories, supporting our finding that taxi drivers are more likely to follow the shortest-distance routes as roadways' hierarchical level increases. Moreover, the category of variably low NonSDR ratios has a higher percentage of state roads located in suburban areas than the category of persistently low NonSDR ratios, also true for variably low NonGTR ratios versus persistently low NonGTR ratios. When comparing between Figure 4.4 (b) and (d) and Figure 4.5 (b) and (d), we can see the pattern that suburban high-hierarchy roads have higher temporal variation in taxi drivers' choices. That is, taxi drivers' preference or avoidance of high-hierarchy roads is more time-dependent in the suburban area than in the urban area. The reason might be that with fewer optional low-hierarchy roads in the suburban area, taxi drivers tend to detour to a high-hierarchy road at some time during a day when its traffic speed is much higher than the local street on the shortest-distance route. However, when the high-hierarchy roads in the suburban area have a large traffic volume at some time during a day, the local streets in the same area usually have good traffic conditions. Thus, taxi drivers tend to detour to those local streets to save time.

4.4.2 Travel Distance and Taxi Drivers' Road Class Preference

Ramaekers et al. (2013) note that trip distance has a significant influence on deviations from the shortest paths. To answer other questions, such as whether for longer trips taxi drivers prefer highways as alternative roads to the shortest-distance routes, this section explores the aggregate patterns of taxi drivers' road class preference in actual routes compared to the shortest-distance routes as travel distance changes. Here we used a trip-based method that is different from the one in the subsection above. For the over 440,000 trips in a day, we assigned each trip, based on its travel distance, to one of the 49 groups with travel distance

intervals of 0-1 km, 1-2 km, 2-3 km, …, and 48-49 km (x axis in Figure 4.7). We used the shortest distance between origin and destination as a reference to define the travel distance on the road network. Smaller travel distance intervals (e.g., intervals of 0-0.5 km, 0.5-1 km, …, and 38.5-49 km) were tested, and there was not much difference in the resulting patterns. For trips in travel distance interval *i*, we also counted the frequency with which road segments of functional class *k* are part of their shortest-distance routes ($SDR\_N^k(i)$), and the frequency with which road segments of functional class *k* are part of their actual routes ($GTR\_N^k(i)$). Next, we calculated the percentage share of functional class *k* among these trips' shortest-distance routes ($SDR\_P^k(i) = SDR\_N^k(i) / \sum_k SDR\_N^k(i)$) and among these trips' actual routes ($GTR\_P^k(i) = GTR\_N^k(i) / \sum_k GTR\_N^k(i)$), respectively. Finally, the difference of functional class *k*'s share in the actual routes vs. in the shortest-distance routes of trips with a travel distance in interval *i* ($Dif\_P^k(i) = GTR\_P^k(i) - SDR\_P^k(i)$) was plotted in Figure 4.7.

The sign of the y axis indicates preference (if positive) or avoidance (if negative) of a functional class in actual routes compared to the shortest-distance routes, and the absolute value indicates the magnitude of preference or avoidance (e.g., a higher positive number indicates a higher degree of preference). For example, point A (46-47, 0.34) and point B (38-39, 0.07) in Figure 4.7 indicate that trips with a travel distance of 46-47 km may prefer state roads more than trips with a travel distance of 38-39 km. Points C and point D indicate that trips with a travel distance of 46-47 km avoid city expressways more than trips with a travel distance of 31-32 km. Figure 4.7 illustrates the following aggregate patterns when comparing actual routes to the shortest-distance routes:

(1) When travel distance is short (<5 km), taxi drivers do not tend to prefer any particular road class. Thus, road functional class might not be an important factor for short trips.

(2) When travel distance is over 30 km, taxi drivers start to consider more state roads in their actual routes. This preference is greatly increased when travel distance is beyond 40 km. As shown in Figure 4.1, state road class contains a beltway surrounding the urban area and connecting the three towns separated by two rivers in Wuhan. A trip of 30+ km usually involves crossing the Yangtze River, thus needing to take one of the three bridges. The state road class has one bridge on its route but is located on the outskirt. Although it has better traffic conditions than the one located in the heart of the urban area, drivers must detour a lot to take advantage of this option. The result indicates that taxi drivers tend to choose the beltway sacrificing distance once the travel distance is long enough.

(3) When travel distance is beyond 21 km, taxi drivers tend to reduce the percentage of city expressways in their routes. This reduction could be related to the fact that those expressways are the most important urban arterials and have

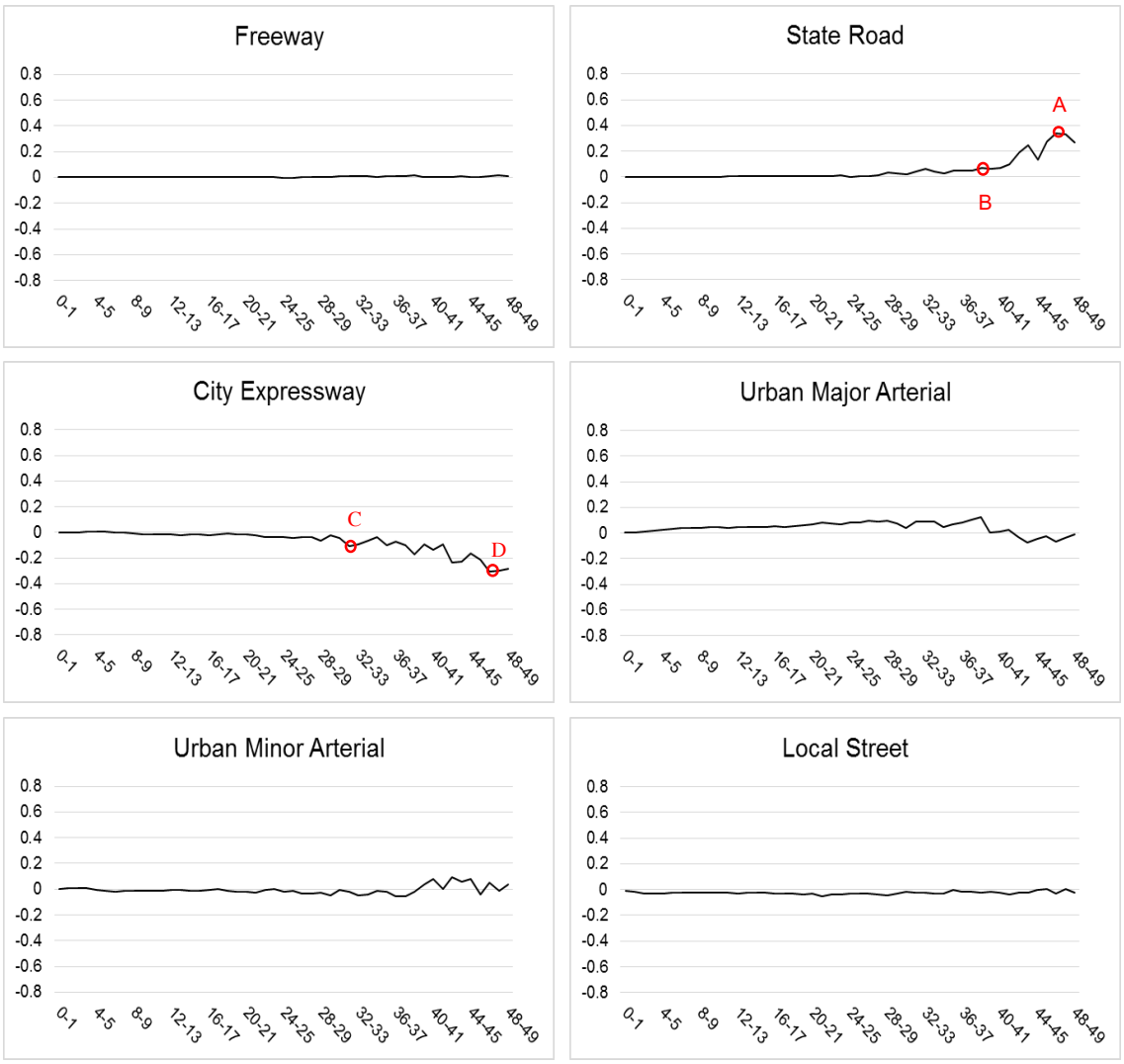


Figure 4.7 Travel distance and road class preference (x axis: travel distance interval in km; y axis: difference of a functional class' share in the actual routes vs. in the shortest-distance routes).

the city's highest traffic volumes. Congestions frequently happen during rush hours on those roads. A longer travel could save more time than a short travel if drivers avoid such roads during rush hours. In fact, starting from 30 km, drivers tend to largely avoid them. This avoidance might be related to the situation, discussed in the previous paragraph, that drivers are likely to take a longer route to get on a beltway and avoid the critical bridge in the expressway class. The 21 km and 30 km could be the thresholds for local and global detours, respectively.

(4) When travel distance is between 5 km and 40 km, drivers tend to slightly prefer urban major arterials as opposed to slightly deviating from urban minor arterials. The major arterials have higher speed limits than minor arterials, but they also have high traffic volumes in rush hours. Taxi drivers may have adjusted their strategies accordingly, not making the overall preference obvious.

(5) Regardless of travel distance, there appears to be no noticeable preference regarding freeways and local streets. Figure 4.1 shows that all freeways are located outside urban areas with fewer optional roads. This geographic characteristic results in high rates of conformance to the shortest-distance routes on freeways. However, local streets are quite different. Our finding, as noted in the last subsection, indicates frequent deviations from the shortest-distance routes in areas with a high density of local streets. The little difference in percentage of local streets can further explain that those deviations are within the same functional class in a local area (i.e., experienced drivers frequently detour from congested local streets to nearby non-congested local streets rather than to other functional classes' roadways).

4.4.3 Deviation Patterns Under Different Situations

Subsection 4.4.1 examines how taxi drivers deviate from the shortest-distance routes on different classes of roads and at different times. It does not differentiate choice behavior between long trips vs. short trips. Subsection 4.4.2 examines taxi drivers' road class preference as travel distance increases, but doesn't consider temporal variation. This section considers the influence of road class, travel distance, and time of day simultaneously to explore route choice patterns under different situations. We first define four situations by combining travel distance and time of day as follows: *long trip during rush hours*, *long trip during off-rush hours*, *short trip during rush hours*, and *short trip during off-rush hours*. Based on Wuhan's public transit administration's official definition of peak hours, we define our rush hours to be 7:00 AM – 9:00 AM and 5:00 PM – 7:00 PM. This definition is consistent with the time windows having the urban roads' lowest average traffic speeds (see Figure 4.8). Each road segment's traffic speed in each time window is roughly estimated by calculating the mean of GPS speeds of the taxis (with passengers) that passed through that road segment in that time window. We used the average

travel distance (5 km) of all taxi trips with passengers to separate long trips from short trips.

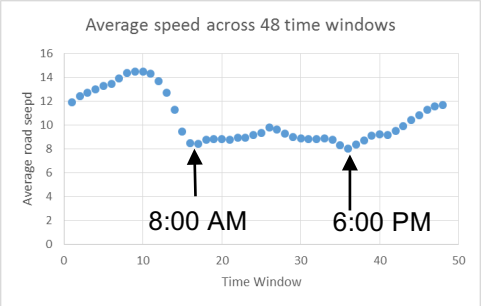


Figure 4.8 Temporal variation of city roads' average speed.

In each of the four situations, we calculated NonSDR Ratio and NonGTR Ratio for each road segment using taxi trips under that situation and mapped them in Figure 4.9(a-d) and Figure 4.10(a-d), respectively. For example, a road segment with a NonGTR Ratio of 0.7 in the situation of long trip during rush hours means that 70% of long trips deviate from this road during rush hours although it is on their shortest-distance routes. To examine how taxi drivers' avoidance (NonGTR Ratio) or preference (NonSDR Ratio) of a road segment changes in different situations, we calculated each road segment's NonSDR Ratio difference and NonGTR Ratio difference in different situations and mapped the road segments with noticeable changes (i.e., with the difference's absolute value greater than 0.1) as shown in Figure 4.9(e-h) and Figure 4.10(e-h).

We found that only about 10% of roads are noticeably different in deviation rate between rush hours vs. off-rush hours for both long and short trips, meaning that for most road segments, taxi drivers' preference for or avoidance of them is not influenced much by rush hours. This lack of influence is probably because during rush hours, the traffic conditions on most road segments become worse at the same time. Taxi drivers may use routing strategies similar to the ones they use during off-rush hours. Among the 10% of roads, the bridge (A in Figure 4.9(e)) to the north of the urban area is more frequently used during rush hours on long trips as an alternative to the main bridge located in the heart of Wuhan. However, the route choice difference between long and short trips is much more noticeable (Figure 4.9 (g-h) and Figure 4.10(g-h)). Over 70% of roads differ in deviation rate above 0.1. The road segments in red indicates a higher deviation rate in long trips vs. short trips, while the blue indicates a lower deviation rate in long compared to short trips. The dominance of red segments indicates that detours are much likely in long vs. short trips, which is consistent between rush hours and off-rush hours.

To further examine road functional class's influence, we plotted the cumulative distribution of Non-SDR ratios and Non-GTR ratios in each road class and

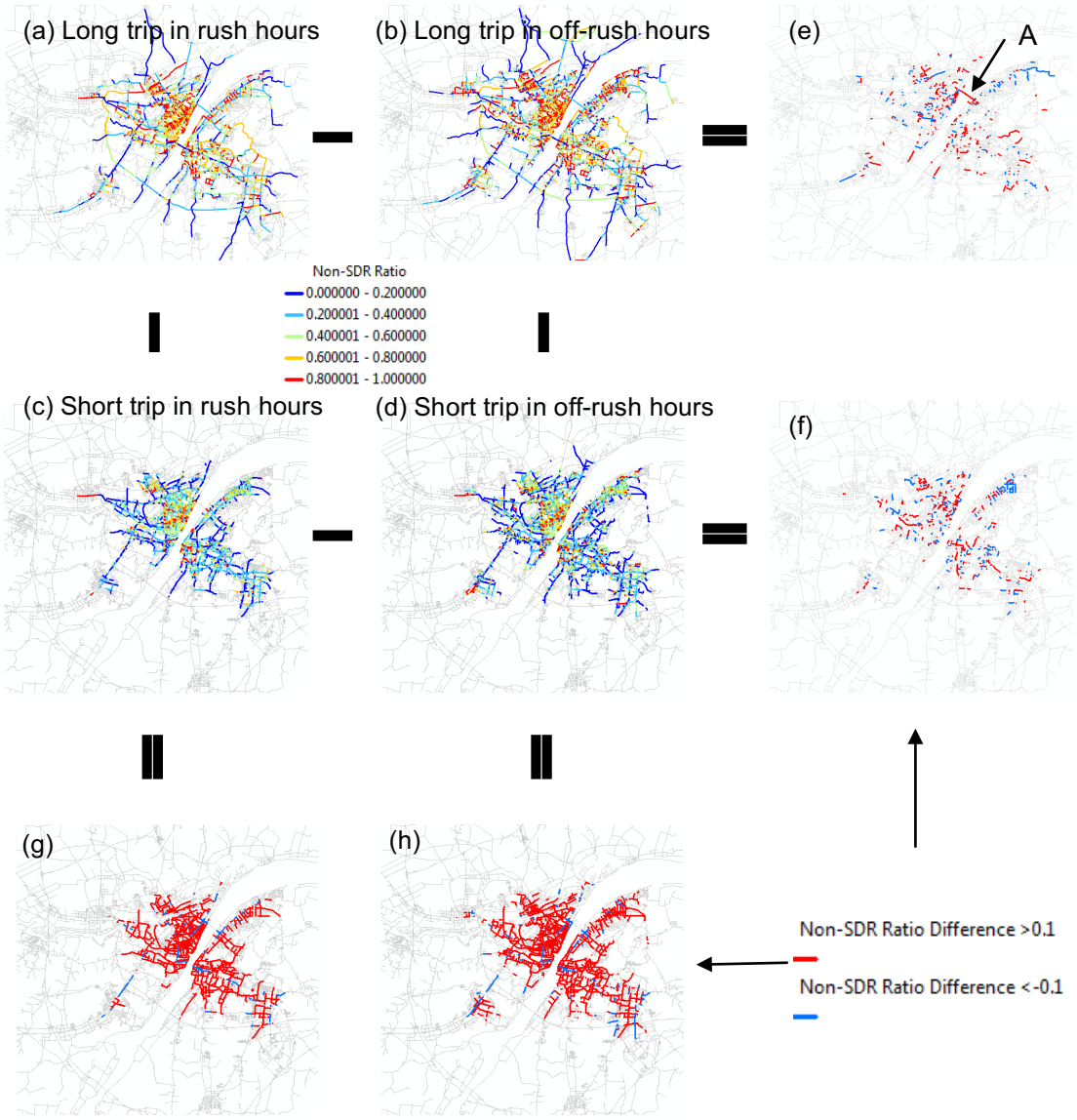


Figure 4.9 (a-d) Non-SDR ratio's spatial distribution in four situations. (e-h) Roads with a noticeable difference in the Non-SDR ratio between (a) and (b), (c) and (d), (a) and (c), (b) and (d).

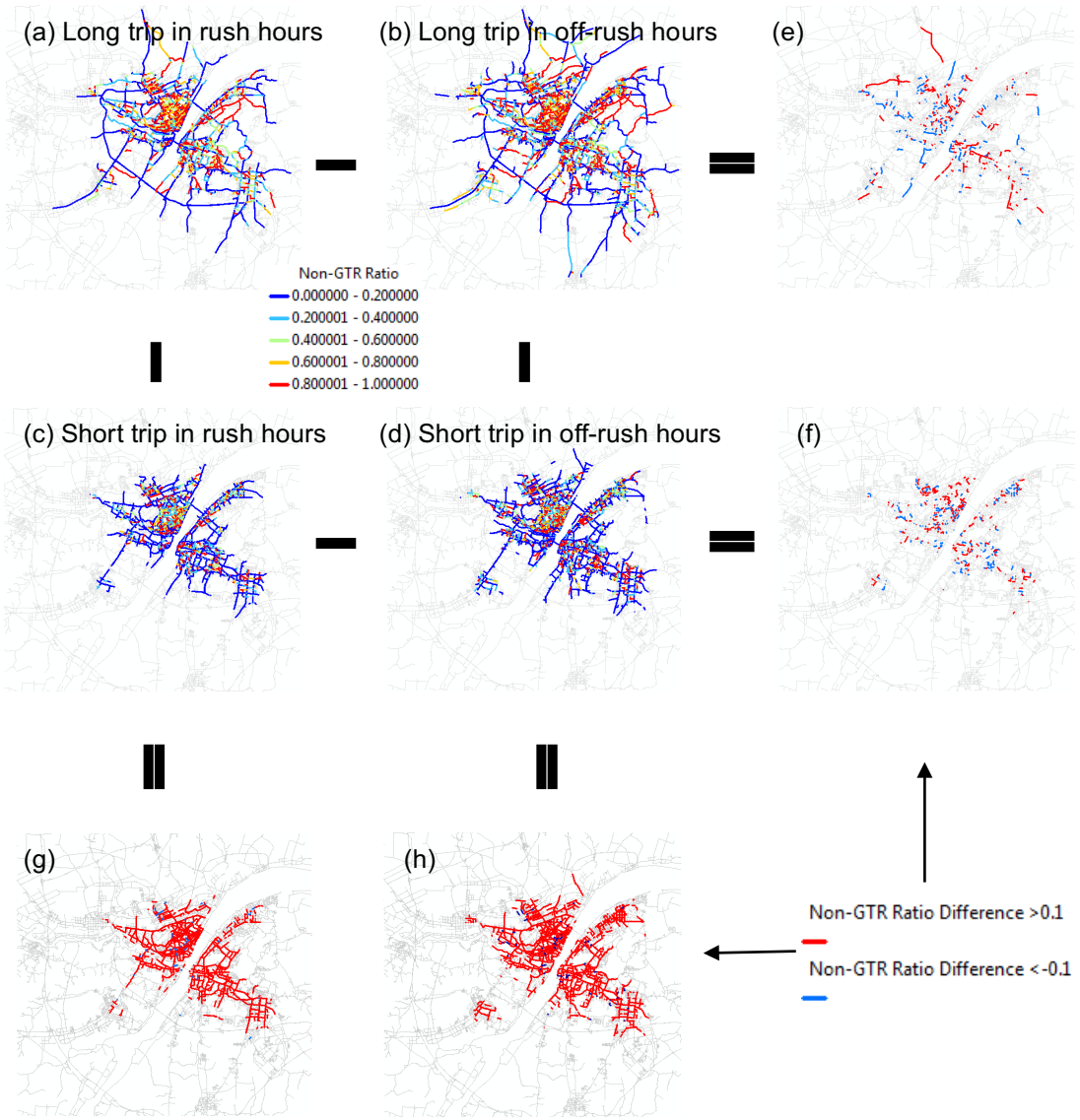


Figure 4.10 (a-d) Non-GTR ratio's spatial distribution in four situations. (e-h) Roads with a noticeable difference in the Non-GTR ratio between (a) and (b), (c) and (d), (a) and (c), (b) and (d).

displayed them by functional class (Figure 4.11) and situation (Figure 4.12). Figures 4.11 and 4.12 have the same content but have different group layouts for easier comparison. In both figures, each line represents the cumulative percentage of Non-SDR ratios or Non-GTR ratios in each road class. A lower line indicates that the road class has more roads with a higher deviation rate than other classes. In other words, deviations are more likely with roads in that class.

Based on Figure 4.11, we find that on roads of any class and in both rush and off-rush hours, long trips are more likely to deviate from the shortest-distance routes than short trips. This finding is consistent with our intuition and conclusions in the literature. However, the literature does not indicate whether road class and rush hours may influence the conclusion. A more interesting finding is that on roads of any class, the short trips' deviation rate seems not influenced by rush hours (see overlapping close blue and yellow lines in Figure 4.11), while on local streets, long trips appear to have more frequent deviations during rush hours than during off-rush hours (see red and green lines in Figure 4.11(a5) and (b5)). Figure 4.11 shows similar patterns across different situations with low-hierarchy roads having high deviation rates. This finding confirms that taxi drivers are more likely to deviate from the shortest-distance routes on low-hierarchy roads in areas with a high-density network than on urban primary roads; this finding is not influenced by travel distance or rush hours. Therefore, deviation from the shortest-distance route is influenced more by road functional class and travel distance than by rush hours. Long trips seem to be influenced more by rush hours than short trips when in local-street areas. This finding is probably because taxi drivers of long trips are more motivated to make detours during rush hours in order to reduce time on a single trip.

4.5 Conclusions

Although the literature has shown a long history of understanding and representing drivers' routing behavior, detailed information about route choice decisions under various situations is still scant. This paper uses a large number of taxi GPS trajectories in a city to examine the spatiotemporal patterns of taxi drivers' deviation from the shortest-distance route, a commonly used assumption in many transportation models and navigation systems.

This paper has proposed two indices for each road segment in a city. These indices contain information about the empirical priority of each roadway to be used in future applications (e.g., suggesting directions and making traffic policies). For those roads that taxi drivers consistently prefer or avoid regardless of time of day, future work can investigate the explanatory variables that this paper has not covered because of lack of road segments' attributes (e.g., speed limit, road width, number of lanes, intersection delay, facility, and urban function of the area along the road). When the factors explaining drivers' consistent avoidance of some roads

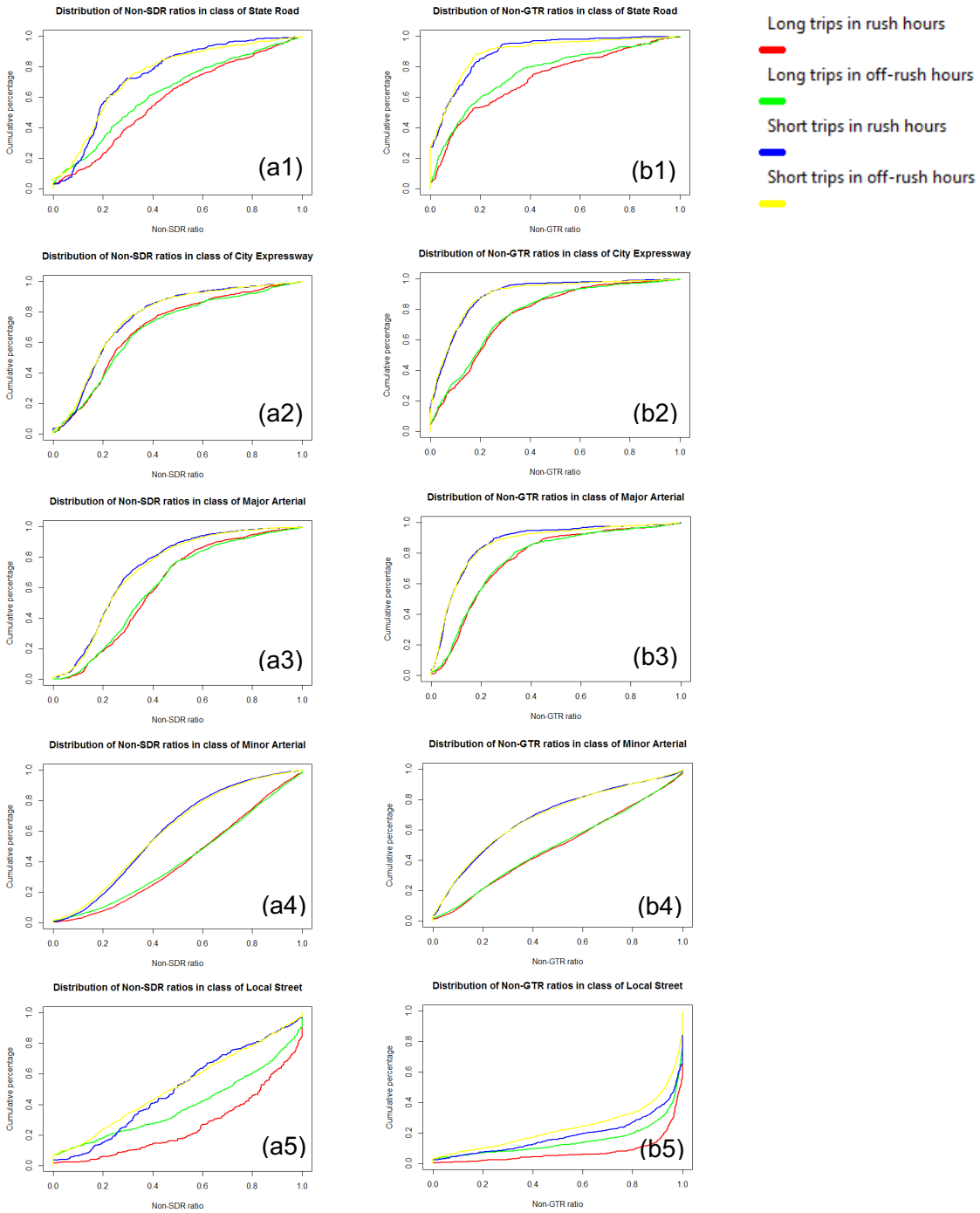


Figure 4.11 Cumulative distribution of (a1-a5) Non-SDR ratios and (b1-b5) Non-GTR ratios in the road classes of (a1,b1) state road; (a2, b2) city expressway; (a3, b3) urban major arterial; (a4, b4) urban minor arterial; and (a5, b5) local street, by different situations.

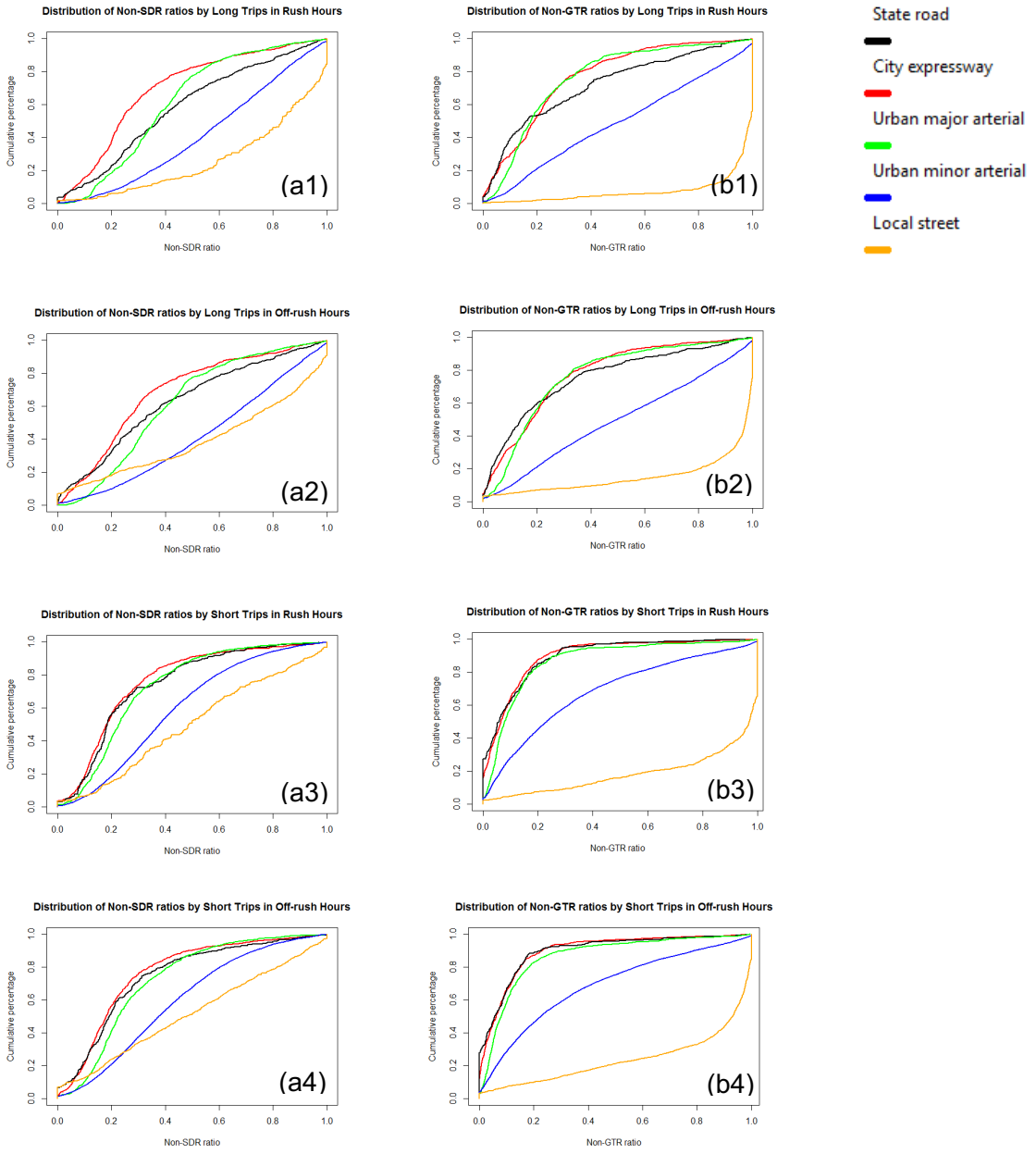


Figure 4.12 Distribution of (a1-a4) Non-SDR ratios and (b1-b4) Non-GTR ratios in different road classes in the following situations: (a1, b1) long trip in rush hours; (a2, b2) long trip in off-rush hours; (a3, b3) short trip in rush hours; (a4, b4) short trip in off-rush hours.

are determined, improvements can be made accordingly to increase these roads' utilization.

Drivers' route-choice behavior observed in this study is more complex than what most transportation models have assumed. Unlike what the research based on traditional sample and survey data suggests, this study found that taxi drivers tend to follow the shortest-distance routes on roads of higher hierarchy and deviate from the shortest-distance routes in areas with a high density of local streets. This conclusion does not suggest that taxi drivers would like to detour to a higher-level road. In fact, they consistently tend not to detour to a primary road if it is not on the shortest-distance route. However, once a primary road is on the shortest-distance route, taxi drivers tend to stay on it regardless of time-varying traffic conditions.

This study also found that taxi drivers' deviation from the shortest-distance route is influenced more by road functional class and travel distance than by urban rush hours. Taxi drivers are more likely to deviate from the shortest-distance routes for longer trips. For short trips (<5km), they do not show preference for any particular road class. However, when travel distance is beyond 30 km and the trip involves river crossing, taxi drivers tend to detour to a city beltway in the outskirt and avoid a busy city expressway in the city center. Moreover, in areas with a high density of local streets, taxi drivers of long trips (>5km) tend to make more detours during rush hours than during off-rush hours. In contrast, for short trips, they do not deviate more during rush hours than during off-rush hours. All the above-mentioned patterns can be used for future simulations of route choice decisions in different situations.

Taxi drivers' preference or avoidance of most urban roads is not influenced much by rush hours. The reason might be that being constrained by most roadways' reduced traffic speeds during rush hours, taxi drivers are likely to use routing strategies similar to the ones used during off-rush hours. This phenomenon could be related to the characteristics of the underlying transportation and urban systems. Moreover, the explored aggregate patterns of taxi drivers' deviations from the shortest-distance routes are similar on weekdays and the weekend. This finding may indicate that although dynamic traffic conditions can influence a driver's route choice decision at the individual level, the aggregate deviation patterns at the systemic level are relatively stable. This conclusion suggests the validity of using a preprocessed network system with relatively static empirical information for modeling route choice decisions. This approach can help mitigate the computational challenge in dealing with dynamic traffic conditions.

Although this study's findings are based only on taxi drivers' trajectories, the effective methods proposed in this paper can be easily applied to non-taxi drivers' GPS trajectories (e.g., commuter drivers). Future work can use these methods to explore the route-choice differences among different driver groups and examine if

taxi drivers make better route decisions than other drivers. Since this study's data was collected from only one city, whether the uncovered patterns are universal or specific to Wuhan is unknown. Future work can apply these methods to taxi GPS tracking data collected from other cities and examine the differences in deviation patterns among different cities. The comparisons among cities can help explore how transportation and urban systems influence drivers' route-choice behavior.

References

Abdel-Aty, M.A., Kitamura, R., and Jovanis, P.P., 1997. Using stated preference data for studying the effect of advanced traffic information on drivers' route choice. *Transportation Research Part C: Emerging Technologies*, 5(1), 39-50.

Bekhor, S., Ben-Akiva, M.E., and Ramming, M.S., 2006. Evaluation of choice set generation algorithms for route choice models. *Annals of Operations Research*, 144(1), 235-247.

Ben-Akiva, M. and Bierlaire, M., 1999. Discrete choice methods and their applications to short term travel decisions. *In*: R.W. Hall, ed. *Handbook of transportation science*. New York: Springer, 5–33.

Cascetta, E., et al., 1996. A modified logit route choice model overcoming path overlapping problems: Specification and some calibration results for interurban networks. *In*: J.B. Lesort, ed. *Proceedings of the 13th international symposium on transportation and traffic theory*. Lyon, France: Pergamon Press, 697–711.

Chen, T.Y., Chang, H.L., and Tzeng, G.H., 2001. Using a weight-assessing model to identify route choice criteria and information effects. *Transportation Research Part A: Policy and Practice*, 35(3), 197-224.

Daganzo, C.F. and Sheffi, Y., 1977. On stochastic models of traffic assignment. *Transportation Science*, 11(3), 253-274.

Dalton, R.C., 2003. The Secret is to follow your nose route path selection and angularity. *Environment and Behavior*, 35(1), 107-131.

Dia, H., 2002. An agent-based approach to modelling driver route choice behaviour under the influence of real-time information. *Transportation Research Part C: Emerging Technologies*, 10(5), 331-349.

Dial, R.B., 1971. A probabilistic multipath traffic assignment model which obviates path enumeration. *Transportation Research*, 5(2), 83-111.

Ehmke, J.F., Meisel, S., and Mattfeld, D.C., 2012. Floating car based travel times for city logistics. *Transportation Research Part C: Emerging Technologies*, 21(1), 338-352.

Frejinger, E. and Bierlaire, M. 2007. Capturing correlation with subnetworks in route choice models. *Transportation Research Part B: Methodological*, 41(3), 363-378.

Giannotti, F., et al., 2011. Unveiling the complexity of human mobility by querying and mining massive trajectory data. *The VLDB Journal—The International Journal on Very Large Data Bases*, 20(5), 695-719.

Jan, O., Horowitz, A.J., and Peng, Z.R., 2000. Using global positioning system data to understand variations in path choice. *Transportation Research Record: Journal of the Transportation Research Board*, 1725, 37-44.

Levinson, D. and Zhu, S., 2013. A portfolio theory of route choice. *Transportation Research Part C: Emerging Technologies*, 35, 232-243.

Li, H., Guensler, R., and Ogle, J., 2005. Analysis of morning commute route choice patterns using global positioning system-based vehicle activity data.

Transportation Research Record: Journal of the Transportation Research Board, 1926, 162-170.

Li, H., 2004. *Investigating morning commute route choice behavior using global positioning systems and multi-day travel data*. Thesis (PhD), Georgia Institute of Technology.

Li, Q., et al., 2011. Path-finding through flexible hierarchical road networks: An experiential approach using taxi trajectory data. *International Journal of Applied Earth Observation and Geoinformation*, 13(1), 110-119.

Liu, L., Andris, C., and Ratti, C., 2010. Uncovering cabdrivers' behavior patterns from their digital traces. *Computers, Environment and Urban Systems*, 34(6), 541-548.

Liu, Y., et al., 2012. Urban land uses and traffic 'source-sink areas': Evidence from GPS-enabled taxi data in Shanghai. *Landscape and Urban Planning*, 106(1), 73-87.

Papinski, D. and Scott, D.M., 2011. A GIS-based toolkit for route choice analysis. *Journal of Transport Geography*, 19(3), 434-442.

Papinski, D. and Scott, D.M., 2013. Route choice efficiency: an investigation of home-to-work trips using GPS data. *Environment and Planning A*, 45(2), 263-275.

Papinski, D., Scott, D.M., and Doherty, S.T., 2009. Exploring the route choice decision-making process: A comparison of planned and observed routes obtained using person-based GPS. *Transportation Research part F: Traffic Psychology and Behaviour*, 12(4), 347-358.

Parkany, E., et al., 2006. Modeling stated and revealed route choice: consideration of consistency, diversion, and attitudinal variables. *Transportation Research Record: Journal of the Transportation Research Board*, 1985, 29-39.

Prato, C.G., 2009. Route choice modeling: past, present and future research directions. *Journal of Choice Modelling*, 2(1), 65-100.

Prato, C.G., Bekhor, S., and Pronello, C., 2012. Latent variables and route choice behavior. *Transportation*, 39(2), 299-319.

Ramaekers, K., et al., 2013. Modelling route choice decisions of car travellers using combined GPS and diary data. *Networks and Spatial Economics*, 13(3), 351-372.

Ramming, M.S., 2002. *Network knowledge and route choice*. Thesis(PhD). Massachusetts Institute of Technology.

Rossetti, R.J., et al., 2000. An agent-based framework for the assessment of drivers' decision-making. *In*: *Intelligent Transportation Systems, 2000. Proceedings. 2000 IEEE*. IEEE, 387-392.

Spissu, E., Meloni, I., and Sanjust, B., 2011. Behavioral analysis of choice of daily route with data from global positioning system. *Transportation Research Record: Journal of the Transportation Research Board*, 2230, 96-103.

Tawfik, A.M., Rakha, H.A., and Miller, S.D., 2010. Driver route choice behavior: Experiences, perceptions, and choices. *In*: *Intelligent Vehicles Symposium (IV), 2010 IEEE*. IEEE, 1195-1200.

Wardrop, J.G., 1952. Some theoretical aspects of road traffic research. *Proceedings of the Institute of Civil Engineers*, 1(3), 325–378.

Zhang, L. and Levinson, D., 2008. Determinants of route choice and value of traveler information: a field experiment. *Transportation Research Record: Journal of the Transportation Research Board*, 2086, 81-92.

Zhu, S., 2010. *The roads taken: theory and evidence on route choice in the wake of the I-35 W Mississippi River Bridge Collapse and Reconstruction*. Thesis (PhD). University of Minnesota.

Zhu, S. and Levinson, D., 2010. *Do people use the shortest path? An empirical test of Wardrop's first principle*. Available from: http://nexus.umn.edu/papers/ShortestPath.pdf [Accessed December 1, 2014].

Chapter 5

Conclusions

5.1 Summary

Representing the spatial extent of places and studying route-choice behavior are two of big data's many applications that reflect human knowledge and behavior as well as improve GIS tools and services. Chapters 2 and 3 focus on handling challenges when estimating the spatial extent of places based on Flickr geotagged photos. Chapter 4 explores spatiotemporal patterns of taxi drivers' deviations from the shortest-distance routes. This final chapter summarizes this dissertation's results, findings and contributions by answering the research questions identified in Chapter 1.

5.1.1 Estimation of Spatial Extent of Places Based on Flickr Geotagged Photos

1. To what degree can biased representation and no quality assurance influence the estimation of place extents based on geotagged photos? Can these problems be overcome? What are some effective approaches to overcoming them?

Some existing research (e.g., Hollenstein and Purves 2010) successfully used geotagged photos to identify the location of places, but a concern remains about whether the derived vague spatial extent can be distorted because of VGI data bias such as the uneven spatial distribution of geotagged photos. The correlation analysis between California's derived vague extent and the study area's popularity density surface discussed in Chapter 2 confirms this concern: popular locations are significantly overestimated on a place's derived spatial extent; the biased representation in Flickr photos' spatial coverage noticeably influences the estimation of place extents using geotagged photos.

To overcome this problem, Chapter 2 argues that each Flickr photo's representativeness should be different. For photos located in an unpopular location, considering the location's disadvantage in photo availability, the photos' importance should be increased. By modeling the representativeness of each photo to be inversely proportional to its location's popularity, Chapter 2 mitigates the influence of biased representation in Flickr photos' spatial coverage, thus deriving a better representation of a place's spatial extent than the traditional KDE method that cannot address the data bias problem.

In Chapter 2 an outlier removal process is proposed based on Delaunay triangulation to handle the issue of no quality assurance. For most of the study cases, the process successfully removes the outliers and keeps those points located within the target place but in low-density areas. Chapter 3 makes an improvement in this outlier removal process that maximizes the major cluster's likelihood ratio instead of setting a static upper limit of the outliers' percentage. Although the outliers and photo availability's spatial heterogeneity are two major challenges in estimating a place's extent based on Flickr photos, their impact can

be reduced to the minimum by applying proper methods as demonstrated in Chapter 2.

Regarding places with disjoint extents, the uneven distribution of Flickr photos and clustering outliers pose a challenge in determining a place's disjoint extents. To distinguish those outlier clusters from the real clusters representing a place's extents, a scan statistic approach is applied in Chapter 3 to detect bursts of photos with a target place name. The bursts are considered very likely to be the locations of place extents. Although the scan statistic method is rarely used in VGI data, the results show good estimations regarding the number of disjoint extents. Dense outlier clusters located in very popular areas are successfully removed, while loose real clusters located in unpopular areas are correctly identified as the locations of a place's disjoint extents. The good results suggest that random user-generated errors' impact on estimating disjoint extents can be successfully eliminated by comparing to the simulations of a random process.

However, it is challenging to deal with an issue caused by the fact that only a small proportion of Flickr users upload the majority of Flickr photos. If a frequent user contributes many photos with incorrect information in an area where very few other users have uploaded photos, the estimate of this area would be biased toward the incorrect information. The influence of the frequent user's errors can be small at locations where an adequate number of other users have made accurate contributions. However, in areas with few contributors, more place-related information and criteria other than the photos' point distribution should be considered to mitigate the impact of photos with incorrect information and to improve the approximation of place extents. Chapter 3 sets a minimum number of contributors for determining if a detected burst can be considered the location of a place's extent. However, a more robust method incorporating some semantic information should be explored in future work to better handle the challenge of users' incorrect contributions in unpopular areas when estimating a place's disjoint extents.

2. How effective are the geotagged photos for deriving vague spatial extents of places? To what degree can these extents be correctly approximated?

For many places in Chapter 2, the derived RW-KDE surfaces show high probabilities near reference boundaries' centers (see Figure 2.3) and produce lower probabilities near the boundary lines. This representation reflects the vagueness of a place's boundary that central locations usually have higher probabilities of being the target place than the locations near the boundary line. Table 5.1 shows the accuracies of the crisp boundaries extracted from the derived vague extents estimated by the RW-KDE method vs. the traditional KDE method with a smoothing bandwidth of 1/2. To be comparable, an extracted crisp boundary's size is chosen to be the same as that of the reference boundary. The

high accuracies (mostly above 0.7) of the crisp boundaries extracted from RW-KDE surfaces indicate good estimations of place extents using geotagged photos.

Table 5.1 Accuracies of the extracted crisp boundaries.

	Manhattan Chinatown	San Francisco Chinatown	Nashville	Philadelphia
RW-KDE	0.82	0.73	0.71	0.64
KDE	0.71	0.64	0.65	0.49
	Rocky Mountain NP	Smoky mountains NP	California	Utah
RW-KDE	0.76	0.70	0.70	0.78
KDE	0.70	0.54	0.48	0.63

For the three places with disjoint extents in Chapter 3, the proposed method also successfully determines the number of a place's disjoint extents and makes a good approximation of all vague extents. The good estimations in Chapters 2 and 3 based on only Flickr geotagged photos indicate that using such geotagged photos to construct acceptable representations of places' vague extents is feasible where a higher value indicates a higher probability of being the target place.

5.1.2 Spatiotemporal Patterns of Taxi Drivers' Deviations from the Shortest-Distance Routes

1. Where, when, and to what extent do taxi drivers deviate from the shortest-distance routes? What are some effective methods for facilitating spatiotemporal analysis of taxi drivers' deviation patterns?

Chapter 4 proposes two indices to measure each road segment's degree of preference and avoidance by taxi drivers compared to the shortest-distance routes across different time windows. The index values for each road segment and in each time window explicitly convey where, when and to what degree taxi drivers deviated from the shortest routes. The temporal variations of these indices are good indicators of taxi drivers' route changes over time. This chapter proposes an approach to simultaneously considering route choice's spatial and temporal aspects by categorizing road segments into four groups based on the deviation level and the degree of temporal variation. Each category represents a type of routing pattern. Then the spatiotemporal patterns of deviations can be easily observed from each road category's spatial distribution. The method is effective for discovering unknown patterns.

2. Are road functional class, travel distance and urban rush hours related to taxi drivers' deviations from the shortest-distance routes? Which factors are more influential? What unknown patterns can be found from big taxi tracking data? Are the conclusions drawn from such big data consistent with those from traditional survey and sample data?

The spatiotemporal patterns of deviations are found to be related to urban road functional classification. Several studies in the literature (e.g., Li 2004, Ramming 2002, Zhu 2010) suggest that drivers tend to detour to high-hierarchy roads in their actual routes. However, the aggregate patterns of taxi drivers' deviations identified in Chapter 4 suggest that taxi drivers are more likely to follow the shortest-distance routes on urban primary roads than on low-hierarchy roads. Taxi drivers tend not to detour to a primary road if the road is not on the shortest-distance route. However, once the primary road is on the shortest-distance route, it is frequently chosen even during periods with bad traffic conditions. Deviations are more frequently made in areas with a high-density network of local streets than on high-hierarchy roads. These patterns are rarely reported in the literature and are somewhat different from the conclusions of studies based on sampled commuter routes. This finding may suggest a difference in route-choice behavior between taxi drivers and ordinary drivers. A possible explanation could be that the former are more experienced with the complex local street network and tend to choose a better route in the local area rather than simply to detour to a primary road like ordinary drivers might do.

Travel distance is also found to influence taxi drivers' route choices. On any type of road and in both rush and off-rush hours, drivers on long trips are much more likely to deviate from the shortest-distance routes than on short trips. For short trips (<5km), taxi drivers do not have any preference about road class. However, when travel distance is beyond 30 km, taxi drivers start to detour to beltways in the outskirt. This threshold might be determined by the underlying network structure in Wuhan. When travel distance is beyond 21 km, taxi drivers start avoiding urban expressways that are the busiest roads in Wuhan city.

Road functional classification and travel distance appear to be more influential than urban rush hours. On most road segments (about 90%), taxi drivers' preference or avoidance of them changes little during rush hours vs. during off-rush hours. Being constrained by most roadways' reduced traffic speeds during rush hours, taxi drivers may tend to use routing strategies similar to the ones used during off-rush hours.

Chapter 4 also reveals that the explored aggregate patterns of taxi drivers' deviations from the shortest-distance routes are not influenced much by weekdays vs. the weekend. This finding suggests that although the dynamic traffic condition can influence a driver's route choice decision at the individual level, the entire

system's aggregate patterns are relatively stable. The aggregate deviation patterns may be related to the underlying transportation and urban structures. Future work can apply our methods to test other cities with different urban and transportation systems to determine whether urban and transportation structures play an important role in taxi drivers' route-choice patterns.

5.2 Future Work

Based on the three papers' conclusions and limitations, this section identifies the following future research directions.

1. Solution for VGI data bias

Although this dissertation's research shows geotagged photos' capability in estimating a place's spatial extent, Chapter 3 identifies a major challenge VGI data bias causes that could impact determining a place's disjoint extents. The data bias is related to the fact that only a small proportion of Flickr users contributed most of Flickr photos. In an area where very few people uploaded photos, if one user contributed a large number of photos with incorrect information, the estimate about this area would be biased toward the incorrect information. The influence of photos with incorrect information can be mitigated in areas where an adequate number of other users made accurate contributions. Thus, the challenge is to deal with frequent users' uploading large numbers of photos with incorrect information in unpopular areas.

The skepticism about VGI's capability in producing reliable geographic information has led to many studies addressing VGI quality issues (e.g., Goodchild and Li 2012). However, current research has provided limited solutions to the above-mentioned data bias. Robust methods incorporating semantics should be developed to reduce the impact of some frequent users' contributing unreliable contents. A direction could be measuring a photo's reliability based on its contributor's tagging behavior pattern. For example, for those users who tend to use the same set of words to tag their photos taken at different places and with different themes, their photos' reliability should be reduced because batch-tagging behavior is more likely to produce incorrect information.

2. *Conflation of multiple data sources*

A major limitation of the research presented in Chapter 4 is that only taxi trajectories and road functional classification are available and used in the analysis. Information is limited about the urban environments where route choices were made. Route choice decisions are complex and influenced by many factors, such as speed limit, road width, number of lanes, facilities, and urban function of the area along the road. Using only trajectories can hardly uncover the explored patterns' underlying process or explain some counterintuitive patterns. For

example, Chapter 4 identifies some roads that taxi drivers consistently avoided regardless of the time of day. Future research needs both to include more data sources and to search for explanatory variables for this pattern. Once explanatory factors are found, improvements can be made accordingly to increase those roads' utilization.

3. *Pattern reproduction*

Big data are often seen as repositories of unknown knowledge. Patterns are often explored to detect any sign of unknown information. Much pattern-exploring research focuses on datasets collected from one or two cities. It is difficult to know whether the explored patterns are universal or just unique to a city; however, both types of patterns are equally important for understanding the underlying process. For example, the spatiotemporal patterns of taxi drivers' deviations from the shortest-distance routes are similar between weekdays and the weekend. This finding is counterintuitive because the urban dynamics could be quite different between weekdays and the weekend and deviations' spatiotemporal patterns might change accordingly. A possible explanation could be that the aggregate patterns of deviations are influenced more by the underlying urban and transportation systems than by human dynamic movements. Knowing whether other cities have similar patterns can help in understanding the relationship between taxi drivers' route-choice behavior and urban dynamics.

This dissertation does not determine whether other driver groups (e.g., commuters) have similar deviation patterns. Future work can apply the same methods to non-taxi drivers' trajectories to explore how taxi drivers choose their routes differently from other driver groups, examine if taxi drivers make better route decisions than other drivers, and discover route-choice differences among different driver groups.

In conclusion, this dissertation identified several challenges and opportunities when using two different types of big data to explore human knowledge of place and route-choice behavior, respectively. Some effective approaches are presented to deal with data bias and quality issues in Flickr geotagged photos as a type of VGI for harvesting the spatial extent of places. Aggregate patterns regarding drivers' deviation from the shortest-distance routes are explored and discussed to inform transportation applications. This dissertation may help in building big data applications' systematic theory.

References

Goodchild, M.F. and Li, L., 2012. Assuring the quality of volunteered geographic information. *Spatial Statistics*, 1, 110-120.

Hollenstein, L. and Purves, R., 2010. Exploring place through user-generated content: using Flickr tags to describe city cores. *Journal of Spatial Information Science*, 1, 21-48.

Li, H., 2004. *Investigating morning commute route choice behavior using global positioning systems and multi-day travel data*. Thesis (PhD). Georgia Institute of Technology.

Ramming, M.S., 2002. Network knowledge and route choice. Thesis (PhD). Massachusetts Institute of Technology.

Zhu, S., 2010. *The roads taken: theory and evidence on route choice in the wake of the I-35 W Mississippi River Bridge collapse and reconstruction*. Thesis (PhD). University of Minnesota.

Vita

Jiaoli Chen was born in Xiangyang, China, to the parents of Huiyu Chen and Xiangping Li. She attended Huazhong Agricultural University for a Bachelor of Science degree in Geographic Information Science (GIS). With great enthusiasm for GIS, she headed to Wuhan University for graduate education and earned a Master of Science degree in June 2010. After graduation, she came to the U.S. and started her Ph.D. research in January 2011, majoring in geography with research interests covering GIS and transportation. In May 2017, she obtained a Ph.D. degree from the University of Tennessee.